Methods Research Report

Empirical Evidence of Associations Between Trial Quality and Effect Size

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

AHRQ

Agency for Healthcare Research and Quality

Advancing Excellence in Health Care • **www.ahrq.gov**

Methods Research Report

Empirical Evidence of Associations Between Trial Quality and Effect Size

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
http://www.ahrq.gov

Contract No. HHSA 290-2007-10062-I

Prepared by:
Southern California Evidence-based Practice Center
Santa Monica, CA

Investigators:
Susanne Hempel, Ph.D.
Marika J. Suttorp, M.S.
Jeremy N.V. Miles, Ph.D.
Zhen Wang, M.S.
Margaret Maglione, M.P.P.
Sally Morton, Ph.D.
Breanne Johnsen, B.A.
Diane Valentine, J.D.
Paul G. Shekelle, M.D., Ph.D.

AHRQ Publication No. 11-EHC045-EF
June 2011

This report is based on research conducted by the Southern California Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10062-I). The findings and conclusions in this document are those of the author(s), who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decisionmakers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

No investigators have any affiliations or financial involvement (e.g., employment, consultancies, honoraria, stock options, expert testimony, grants or patents received or pending, or royalties) that conflict with material presented in this report.

Suggested Citation:
Hempel S, Suttorp MJ, Miles JNV, Wang Z, Maglione M, Morton S, Johnsen B, Valentine D, Shekelle PG. Empirical Evidence of Associations Between Trial Quality and Effect Sizes. Methods Research Report (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-2007-10062-I). AHRQ Publication No. 11-EHC045-EF. Rockville, MD: Agency for Healthcare Research and Quality. June 2011. Available at: http://effectivehealthcare.ahrq.gov.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers; as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director, Task Order Officer
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgements

We would like to thank the Tufts Medical Center and the Research Triangle Institute – University of North Carolina (RTI-UNC) Evidence-based Practice Center for their collaboration in identifying EPC reports and Kirstin Nyrop and Jonathan V. Todd for quality scoring of selected articles. We also thank Ethan Balk, Nancy Berkman, Isabelle Boutron, Tim Carey, Mark Helfand, David Moher, and Sydne Newberry for comments on an earlier draft of this report.

Empirical Evidence of Associations Between Trial Quality and Effect Sizes

Structured Abstract

Objectives. To examine the empirical evidence for associations between a set of proposed quality criteria and estimates of effect sizes in randomized controlled trials across a variety of clinical fields and to explore variables potentially influencing the association.

Methods. We applied quality criteria to three large datasets of studies included in a variety of meta-analyses covering a wide range of topics and clinical interventions consisting of 216, 165, and 100 trials. We assessed the relationship between quality and effect sizes for 11 individual criteria (randomization sequence, allocation concealment, similar baseline, assessor blinding, care provider blinding, patient blinding, acceptable dropout rate, intention-to-treat analysis, similar cointerventions, acceptable compliance, similar outcome assessment timing) as well as summary scores. Inter-item relationships were explored using psychometric techniques. We investigated moderators and confounders affecting the association between quality and effect sizes across datasets.

Results. Quality levels varied across datasets. Many studies did not report sufficient information to judge methodological quality. Some individual quality features were substantially inter-correlated, but a total score did not show high overall internal consistency (α 0.55 to 0.61). A factor analysis-based model suggested three distinct quality domains. Allocation concealment was consistently associated with slightly smaller treatment effect estimates across all three datasets; other individual criteria results varied. In dataset 1, the 11 individual criteria were consistently associated with lower estimated effect sizes. Dataset 2 showed some unexpected results; for several dimensions, studies meeting quality criteria reported larger effect sizes. Dataset 3 showed some variation across criteria. There was no statistically significant linear association of a summary scale or factor scores with effect sizes. Applying a cutoff of 5 or 6 criteria met (out of 11) differentiated high and low quality studies best. The effect size differences for a cutoff at 5 was -0.20 (95% confidence interval [CI]: -0.34, -0.06) in dataset 1 and the respective ratio of odds ratios in dataset #3 was 0.79 (95% CI: 0.63, 0.95). Associations indicated that low-quality trials tended to overestimate treatment effects. This observation could not be replicated with dataset 2, suggesting the influence of confounders and moderators. The size of the treatment effect, the condition being treated, the type of outcome, and the variance in effect sizes did not sufficiently explain the differential associations between quality and effect sizes but warrant further exploration in explaining variation between datasets.

Conclusions. Effect sizes of individual studies depend on many factors. The conditions where quality features lead to biased effect sizes warrant further exploration.

Contents

Tables

Figures

Appendixes

Executive Summary

Background

Trial design and execution factors are widely believed to be associated with bias. Bias is typically defined as a systematic deviation of an estimate, such as the estimated treatment effect from the true value. More factors have been proposed as associated with bias than have actually been empirically confirmed by systematic examination. There are some conflicting results regarding the association of quality features and effect sizes. Little is known about moderators and confounders that might predict when quality features (or the lack thereof) influence results of research studies.

Objective

The objective of this project was to examine the empirical evidence for associations between a set of proposed quality criteria and estimates of effect sizes in randomized controlled trials using multiple datasets representing a variety of clinical fields and to explore variables potentially influencing the association.

Methods

We applied a set of proposed quality criteria to three large datasets of studies included in a variety of systematic reviews covering a wide range of clinical fields. The first dataset was derived from all Cochrane Back Review Group reviews of nonsurgical treatment for nonspecific low back pain in the Cochrane Library 2005, issue 3; the set included 216 individual trials. For the second dataset we searched prior systematic reviews and meta-analyses conducted by Agency for Healthcare Research and Quality-funded Evidence-based Practice Centers with the goal of assembling a set with a wide range of clinical topics and interventions; this dataset included 165 trials. The third dataset was obtained by replicating a selection of trials used in a published meta-epidemiological study demonstrating associations of quality with the size of treatment effects; this set included 100 trials (79 percent of the original dataset).

The proposed set of 11 quality features comprised the following:

- Generation of the randomization sequence
- Concealment of treatment allocation
- Similarity of baseline values
- Blinding of outcome assessors
- Blinding of care providers
- Blinding of patients
- Acceptable dropout rate and stated reasons for withdrawals
- Intention-to-treat analysis
- Similarity of cointerventions
- Acceptable compliance
- Similar timing of outcome assessment.

In addition we applied the Jadad components and scale, and criteria suggested by Schulz, including allocation concealment, to one of the datasets. The inter-item relationships of the proposed quality criteria were explored using psychometric methods. A multiple indicator multiple cause (MIMIC) factor analysis explored inter-item correlations as well as associations of quality features with reported effect sizes.

We assessed the relationship between quality and effect sizes for individual criteria as well as summary scores. In particular, the use of total quality scores per study with each item adding to a sum score, factor-analytically derived broad quality domains, and the application of different cutoffs for a total quality score was further explored.

We investigated moderators and confounders that affect the association between quality measures and the size of the treatment effect across datasets. In particular, we investigated whether (1) the overall size of the treatment effect of the intervention observed in datasets, (2) the condition being treated, (3) the investigated type of outcome, and (4) the variance in effect sizes across studies moderates or confounds the association between quality and effect sizes.

Results

The average quality levels varied across datasets. Many studies did not report sufficient information to judge the quality of the feature (although quality of reporting increased after the introduction of the Consolidated Standards of Reporting Trials statement). Some individual quality features were substantially intercorrelated, but a total score did not show high overall internal consistency of the 11 quality features (α's = 0.55 to 0.61). A MIMIC factor-analytic model suggested three distinct quality domains; randomization sequence generation and allocation concealment constituted the first factor, the blinding items constituted a second factor, and the third factor was primarily derived from the acceptable dropout rate item.

Allocation concealment was consistently associated with a slightly smaller treatment effect across all three datasets: Effect size differences were −0.08 (95% CI: −0.23, 0.07) in dataset 1 and −0.06 (95% CI: −0.22, 0.11) in dataset 2. The ratio of odds ratios was 0.91 (0.72, 1.14) in the third dataset where only categorical outcome measures were included; hence, we computed odds ratios rather than effect sizes. Other individual criteria results varied across datasets. In dataset 1 the 11 individual quality criteria were consistently associated with a lower effect size, indicating that low-quality studies overestimated treatment effects. Results in dataset 2 showed unexpected results: Higher quality studies reported larger effect sizes in this sample. The third dataset showed some variation across quality criteria.

There was no statistically significant linear association of a summary quality score (derived by equally weighing all 11 quality items) and effect sizes, which would have indicated that the effect size decreased linearly with increased quality. There was also no consistent linear association across datasets for the factor scores.

Applying a cutoff of 5 or 6 quality criteria met (out of a possible 11) differentiated high- and low-quality studies best. Effect size differences were −0.20 in dataset 1. In the third dataset, the ratio of odds ratios were 0.79 (cutoff at 5; 95% CI: 0.63, 0.95) and 0.77 (cutoff at 6; 95% CI: 0.63, 0.99). These associations indicated that low-quality trials tended to overestimate treatment effects. This effect could not be replicated in dataset 2, suggesting the influence of confounders and moderators of the association.

The specific moderators and confounders that were investigated in this report did not sufficiently explain the variation in associations across datasets. When controlling for the mean treatment effect obtained in each included meta-analysis, the differences across datasets in observed associations between quality and effect sizes remained. A stratified analysis for the condition being treated also failed to explain the contrary results observed in dataset 2 compared to the other two datasets; the clinical condition did not appear to confound the underlying association between quality and effect sizes for individual quality criteria, and the interaction effect of condition with total quality score was also not statistically significant. When categorizing the different measures used to show a treatment effect into objective versus more subjective outcomes, the type of outcome did not show statistically significant interaction effects. The variance in effect sizes within datasets varied across the three datasets and may potentially explain differences observed in the association between quality and effect sizes across datasets; this finding should be investigated systematically. Several assumptions can be tested in meta-epidemiological datasets that may help determine when and which quality features lead to biased effect sizes.

Conclusions

The associations between quality features and effect sizes are complex. Effect sizes of individual studies depend on many factors. In two datasets, individual quality items and summary scores of items were associated with differences in effect sizes. This relationship was not found in the remaining dataset. Despite several exploratory analyses, we were not able to explain these differences. The conditions under which quality features and which features lead to biased effect sizes warrant further exploration and factors such as the variance in quality scores and effect sizes will be investigated in a subsequent project.

Background

Trial design and execution factors are widely believed to be associated with bias in randomized controlled trials (RCT). Bias is typically defined as a systematic deviation of an estimate, such as the estimated treatment effect from the true value. A number of individual quality criteria and quality checklists or scales for RCTs have been proposed (see e.g., Moja, Telaro D'Amico, et al. 2005; West, King, Carey, et al., 2002). These cover potential threats to the internal validity of the trial methodology.

Quality checklists typically provide a selection of quality features that are scored individually. Quality scales provide in addition a total quality score, either by summing up individual features giving equal weights to each feature or by putting more emphasis on selected features. Existing quality checklists and scales address the conduct or research methodology of the individual study, so they concern the internal validity of the research study, but they frequently also include other quality aspects of publications. Jadad and colleagues (Jadad, Moore, Carroll, et al., 1996) proposed a scale of 0 to 5 to evaluate RCTs with "low" and "high" internal validity in pain research. The Jadad scale, based on three criteria (randomization, double-blinding, and a description of dropouts), is widely used as a summary quality measure of randomized controlled trials (RCTs) and is one of the few tools where the psychometric properties have been evaluated and are acceptable. However, the Jadad scale has some limitations, e.g., the double blinding criterion is usually reported in fewer than 10 to 20 percent of studies. Many trials involve devices, surgery, or other interventions for which double blinding is either impractical or impossible and the double blinding criterion accounts for 40 percent of the Jadad score. An additional criterion, the concealment of treatment allocation, is not included in the Jadad scale but is widely used in addition to the criteria proposed by Jadad et al. (1996). Verhagen, de Vet, de Bie, et al. (1998) developed a nine-item list of quality items specifically focused on internal validity, using a formal Delphi process of three rounds, which included leading experts from around the world. The 2008 Cochrane handbook (Higgins and Green, 2008) introduced a Risk of Bias tool based on the domains sequence generation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting, and other sources of bias.

More factors have been proposed as related to bias than have actually been confirmed by systematic examination. Only a few researchers have published investigations of the association between selected trial quality and effect sizes obtained in individual trials. It is assumed that the conduct of the research methodology will influence the result that is obtained by the trial. The study methodology appears to distort the true value expected to be shown in the study. Typically, it is assumed that low-quality trials exaggerate treatment effects. Colditz, Miller, and Mosteller (1989) found RCTs to have smaller effect sizes than non-RCTs in studies of surgical therapy and RCTs that are double blind have smaller effect sizes than nonblinded trials of medical therapy. Schulz, Chalmers, Hayes, et al. (1995) assessed 250 trials in 33 meta-analyses and reported that inadequate concealment of allocation accounted for a 41 percent increase in effect sizes. The lack of double blinding showed a 17 percent increase in reported treatment effect. Contrarily, Emerson, Burdick, Hoaglin, et al. (1990) found no relationship between a consensus-developed quality scale (0–100 points) and treatment differences. Balk and colleagues (Balk, Bonis, Moskowitz, et al., 2002) applied 24 existing quality measures and assessed 26 meta-analyses involving 276 RCTs. The analysis focused on four conditions: cardiovascular disease, infectious disease, pediatrics, and surgery. The study found no indication of bias; individual quality measures were not reliably associated with the strength of treatment effect across studies and clinical areas. Moher, Pham, Jones, et al. (1998) used Jadad's scale and Schulz's "concealment of allocation" in a large study assessing 11 meta-analyses including 127 RCTs. All trials were scored and the meta-analyses replicated. Low-quality studies were associated with an increased treatment estimate of 34 percent compared with high-quality trials. Studies with inadequate treatment allocation concealment showed a 37 percent increased effect compared to concealed trials. Juni, Altman, and Egger (2001) have summarized the data from Schulz et al. (1995), Moher et al. (1998), Kjaergard, Villumsen, and Gluud (1999) and Juni, Tallon, Egger, et al. (2000) in a pooled analysis and provide evidence for associations of effect sizes with allocation concealment (ratio of odds ratios [ROR] 0.70; 95% CI: 0.62, 0.80) and double blinding (ROR 0.86; 95% CI: 0.74, 0.99) while the generation of treatment allocation did not have a statistically significant effect across datasets (ROR 0.81; 95% CI: 0.60, 1.09). Pidal, Hrobjartsson, Jorgensen, et al. (2007) outline the potential consequences for meta-analysis conclusions. When only trials with adequate concealment were included in meta-analyses, two-thirds lost statistical significance of the primary result, primarily to loss of power (as a result of a smaller sample size) but also a shift in the point estimate towards a less beneficial effect. These studies provide data on quantifying the risk of bias associated with individual or sets of quality criteria.

The association between quality features and effect sizes may vary across datasets according to factors yet to be explored. Investigating moderators and confounders that may influence the association between quality and effect sizes and that may explain some of the conflicting results shown in the literature is an evolving field. Wood, Egger, Gluud, et al. (2009) used three sets of "meta-epidemiological studies," that is, studies investigating the associations of quality features and effect sizes (Kjaergard, Villumsen & Gluud, 2001; Schulz, et al., 1995; Egger, Juni, Bartlett, Holenstein, and Sterne, 2003). The group investigated whether the nature of the intervention and the type of outcome measures influences the effect of allocation concealment and blinding. They found that studies using subjective outcomes showed exaggerated effect sizes when there was inadequate or unclear allocation concealment or lack of blinding. In studies using objective outcomes such as mortality, the association of quality with trial results was negligible. Differentiating drug interventions and nondrug interventions showed no significant differences on the effect on allocation concealment or blinding.

Recently, quality criteria suggested by the Cochrane Back Review Group (CBRG) were found to be associated with effects sizes in RCTs of interventions for back pain (van Tulder, Suttorp, Morton, et al., 2009). The CBRG Editorial Board developed an 11-item criteria list, based on the 9-item Delphi list (Verhagen et al., 1998) and the 3-item Jadad criteria, for evaluation of internal validity of RCTs (van Tulder, Furlan, Bombardier, et al., 2003). Modifications were made to tailor the criteria list to the expected needs of trials of treatments for back pain. The Delphi list was modified by adding three items that had been eliminated between rounds two and three of the Delphi (items about withdrawals and dropouts, compliance rate, and co-interventions), deleting one Delphi list criterion on specifying eligibility criteria and adding one item about the timing of measurement of outcomes. This 11-item list was then proposed by the CBRG editorial board as the standard measure for assessing quality of controlled trials and has been used in virtually all CBRG reviews. A summary score of 0 to 11, based on the 11-item list, was developed as a measure of overall internal validity. Results of applying this set of criteria on all trials of nonsurgical therapy in CBRG reviews showed consistent effects of the criteria on effect sizes.

We aim to assess the potential usefulness of the set of CBRG quality criteria to other clinical conditions by applying these criteria to large datasets of RCTs covering diverse clinical topics and diverse outcome measures. We examine the empirical evidence of associations between individual quality criteria as well as summary scores. In addition, factors influencing the association between quality and effect sizes are explored.

Methods

This project developed sequentially over time. The original study was part of a project for the Cochrane Back Review Group (CBRG). The additional work was funded by the Agency for Healthcare Research and Quality in steps as results of earlier analyses suggested fruitful areas for testing of new hypotheses.

Quality Criteria

We applied the 11 CBRG Internal Validity criteria (van Tulder et al, 2003) that appeared very promising in the quality scoring of Cochrane back reviews. The items cover established quality criteria (allocation concealment, blinding), as well as criteria for which no evidence on their potential for bias has been investigated or existing studies showed conflicting results.

The individual criteria address the adequacy of the randomization sequence generation, concealment of treatment allocation, baseline similarity of treatment groups, outcome assessor blinding, care provider blinding, patient blinding, adequacy and description of the dropout rate, analysis according to originally assigned group (intention-to-treat analysis), similarity of cointerventions, adequacy of compliance and similar assessment timing across groups. The items and the scoring guideline are shown in Appendix F.

The answer mode employed the following categories: “Yes,” “No,” and “Unclear.” The CBRG offers concrete guidance for each answer category. Assessor blinding for example is scored positively when assessors were either explicitly blinded or the assessor is clearly not aware of the treatment allocation (e.g., in automated test result analysis).

A number of items are topic specific and have to be defined individually. For each topic, a content expert (typically a clinician with trial research experience) was contacted to assist in the selection of baseline comparability variables and to establish reasonable dropout and compliance rates. The baseline comparability assessment requires that topic specific key prognostic predictors of the outcome are specified and the baseline comparability of the treatment groups has to be judged. For interventions that involve considerable patient commitment (e.g., presenting at multiple outpatient appointments) a dropout rate of about 25 percent was considered sufficient, while for other interventions a rate of 10 percent was considered sufficient in order to meet this criterion in the specific clinical area.

In addition, for one of the datasets the Jadad scale (Jadad et al., 1996) and criteria proposed by Schulz et al. (1995), operationalized as in the original publications, was applied. The Jadad scale (0 to 5 points) assesses randomization (0 to 2 points), blinding (0 to 2 points), and withdrawals (0 to 1 point). The applied Schulz criteria were allocation concealment, randomization sequence, analysis of all randomized participants, and double blinding. The items together with the scoring instructions can be found in Appendix F.

Study Pool Selection

This project drew on three different study pools. One was available from previous work for the Cochrane Back Review Group, the project has been described in detail elsewhere (van Tulder et al., 2009). Two datasets were obtained for the purpose of this project only (datasets 2 and 3). First results on the association between quality and effect sizes in dataset 1 have been published previously (van Tulder et al., 2009), all further analyses were prepared for this report only.

Dataset 1: Back Pain Trials

For the CBRG project the quality criteria were originally applied to all CBRG reviews of nonsurgical treatment for nonspecific low back pain present in the Cochrane Library 2005, issue 3. The study set was drawn from 12 reviews (Assendelft, Morton, Yu, et al., 2004; Furlan, van Tulder, Tsukayama, et al, 2005; Furlan, Imamura, Dryden, et al., 2005; Hagen, Hilde, Jamtvedt, et al., 2005; Hayden, van Tulder, Malmivaara, et al., 2005; Henschke, Ostelo, van Tulder, et al., 2005; Heymans, van Tulder, Esmail, et al., 2004; Karjalainen, Malmivaara, van Tulder, et al., 2003; Khadilkar, Odebiyi, Brosseau, et al., 2005; Roelofs, Deyo, Koes, et al., 2005; van Tulder, Touray, Furlan, et al., 2003; van Duijvenbode, Jellema, van Poppel, et al., 2005). Studies reported on pain, function, or other improvement measures. The reviews assessed the effect of acupuncture, back schools, behavioral treatment, exercise therapy, bedrest, lumbar supports, massage, multidisciplinary bio-psycho-social rehabilitation, muscle relaxants, spinal manipulative therapy, and transcutaneous electrical nerve stimulation (TENS) for the treatment of low-back pain. Comparisons were placebo, usual care, or no treatment or comparisons between treatments. The dataset included 216 trials.

Dataset 2: EPC Reports

In the first of two efforts supported by AHRQ, we assembled a second dataset of trials based on Evidence-based Practice Center (EPC) reports. We searched prior systematic reviews and meta-analyses conducted by AHRQ-funded EPCs with the goal of assembling a test set of studies that represented a wide range of clinical topics and interventions. The criteria for selection were that the EPC report had to include a meta-analysis and that the EPC had to be willing to provide us with the data on outcomes, such that we only needed to assess the quality of the included trials. The study set was drawn from 12 evidence reports, the majority were also published as peer review journal articles (Balk, Lichtenstein, Chung, et al., 2006; Balk, Tatsioni, Lichtenstein, et al., 2007; Chapell, Reston, Snyder, et al., 2003; Coulter, Hardy, Shekelle et al., 2003; Donahue, Gartlehner, Jonas, et al., 2007; Hansen, Gartlehner, Webb, et al., 2008; Hardy, Coulter, Morton, et al., 2002; Lo, LaValley, McAlindon, et al., 2003; Shekelle, Morton, Hardy, 2003; Shekelle, Maglione, Bagley, et al., 2007; Shekelle, Morton, Maglione, et al., 2004; Towfigh, Romanova, Weinreb, et al., 2008). The reports addressed a diverse set of topics, pharmacological therapies as well as behavior modification interventions. All studies included in the main meta-analysis of the report were selected; studies included in more than one report entered our analysis only once. The dataset included 165 trials.

The reports addressed pharmaceuticals (orlistat, vitamin E, drugs for arthritis, S-adenosylmethionine, chromium, atypical antipsychotics, omega-3 fatty acids); non-pharmacological studies such as self-monitoring of blood glucose (SMBG), diet and weight loss, chronic disease self-management (CDSM); interventions to manage and treat diabetes (chromium, SMBG, CDSM); complementary and alternative medicine/dietary supplements (vitamin E, chromium, omega-3); as well as mental health topics (Alzheimer's, obsessive-compulsive-disorder [OCD]).

In each of the evidence reports one meta-analysis (in general the analysis with the largest number of trials) was selected and all studies included in that pooled analysis were chosen for the study pool. Only one comparison per study was included. Multiple publications and multiple outcomes were excluded so that each unique study entered the test set only once. In the majority, individual studies compared the intervention to placebo or usual care.

Dataset 3: Published “Pro-bias” Sample

Following the results of the analysis of the EPC reports, we obtained a third dataset of studies. This third dataset was obtained by replicating a selection of trials used by Moher et al. (1998). The dataset was chosen as it has shown evidence of bias for established quality criteria (see Moher et al., 1998) and is designated in this report as “pro-bias.” Since the original publication does not specify exactly which trials and which outcomes were included in this analysis, we replicated the methods described by Moher and colleagues for selection. Two reviewers independently reviewed the 11 meta-analyses chosen by Moher et al. and reconciled their assessment of the primary outcome and the main meta-analysis in the publication. Following the described approach, this designation of the primary outcome was based on the largest number of randomized controlled trials (RCTs) reporting data on that endpoint since many meta-analyses did not identify a primary outcome. Individual trials present in multiple meta-analyses were included only once so that a trial did not enter our analyses more than once. Where multiple comparisons were reported in original articles we included those data chosen in the main analysis of the 11 meta-analyses. We were able to retrieve, quality score, and abstract 100 RCTs of the originally published set (79 percent).

The trials came from meta-analyses on digestive diseases (Marshall and Irvine, 1995; Pace, Maconi, Molteni, et al., 1995; Sutherland, May, and Shaffer, 1993), circulatory diseases (Ramirez-Lasspas and Cipolle, 1988; Lensing, Prins, Davidson, et al., 1995; Loosemore, Chalmers, and Dormandy, 1994), mental health (Mari and Streiner, 1994; Loonen, Peer, and Zwanikken, 1991; Dolan-Mullen, Ramirez, and Groff, 1994), stroke (Counsell Sandercock, 1995) and pregnancy and childbirth (Hughes, Collins, and Vanderkeckhove, 1995).

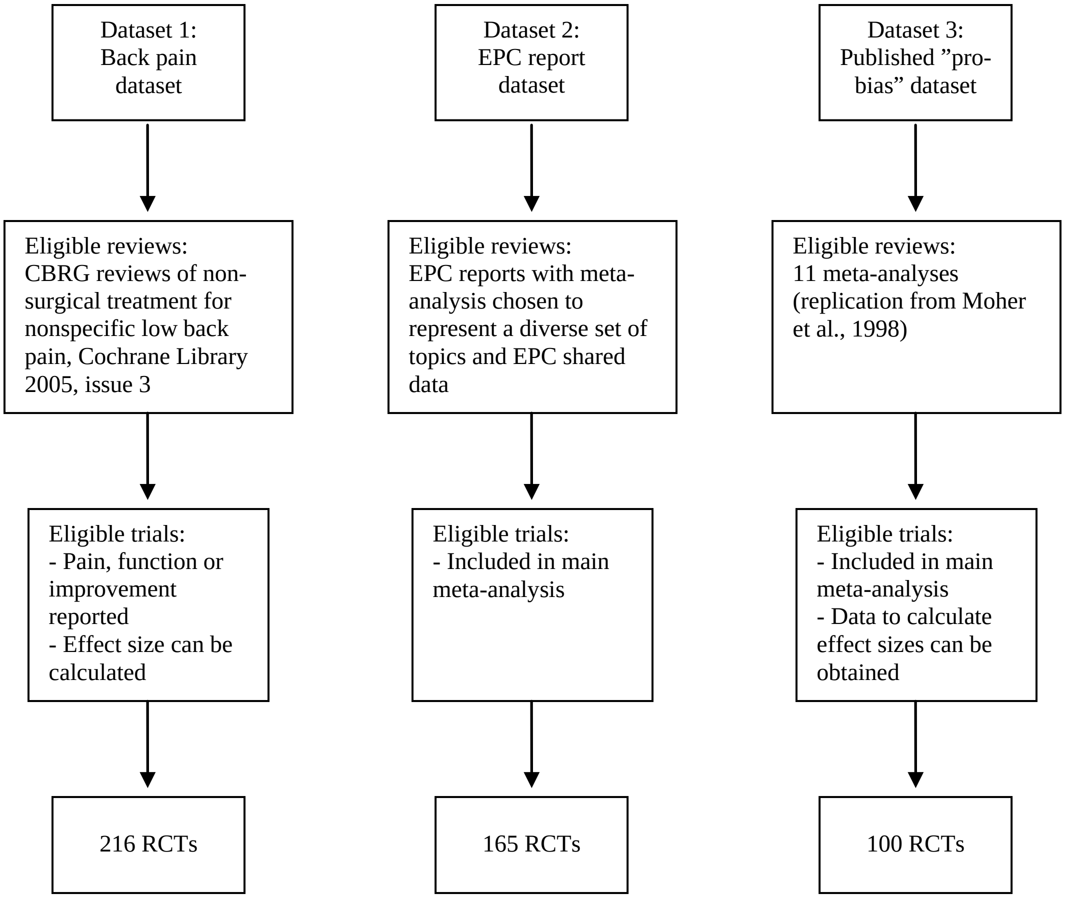
The flow diagram in Figure 1 summarizes the dataset composition.

Procedure

We developed and pilot tested a standardized form to record decisions for the quality criteria. For all datasets, two reviewers independently rated the quality of each study by applying the outlined quality criteria. The reviewers used the full publications to score the studies and were not blinded to authors, journals or other variables. The reviewers were experienced in rating study quality in the context of evidence based medicine and underwent an additional training session for this study. The pair of reviewers reconciled any disagreement through consensus; any remaining disagreements were resolved by discussion in the research team.

The outcomes of the individual RCTs were extracted by a statistician together with measures of dispersion where available and the number of patients in each group. For dataset 1 (back pain) absolute effect sizes were used as this dataset included comparisons between treatment and placebo as well as comparisons between active treatments. For dataset 2 (EPC reports) in order to be able to combine studies within data sets or where possible between datasets, standardized effect sizes were computed for each study. As all studies in dataset 3 (pro-bias) reported dichotomous outcomes, odds-ratios (OR) were calculated. As a quality check, the point estimate and 95 percent confidence interval (CI) of each meta-analysis was calculated and compared to the original meta-analytic result.

Figure 1. Flow diagram summarizing dataset composition


Dataset 1: Back pain dataset
Dataset 2: EPC report dataset
Dataset 3: Published ”pro-bias” dataset
Eligible reviews: CBRG reviews of non-surgical treatment for nonspecific low back pain, Cochrane Library 2005, issue 3
Eligible reviews: EPC reports with meta-analysis chosen to represent a diverse set of topics and EPC shared data
Eligible reviews: 11 meta-analyses (replication from Moher et al., 1998)
Eligible trials: - Pain, function or improvement reported - Effect size can be calculated
Eligible trials: - Included in main meta-analysis
Eligible trials: - Included in main meta-analysis - Data to calculate effect sizes can be obtained
216 RCTs
165 RCTs
100 RCTs

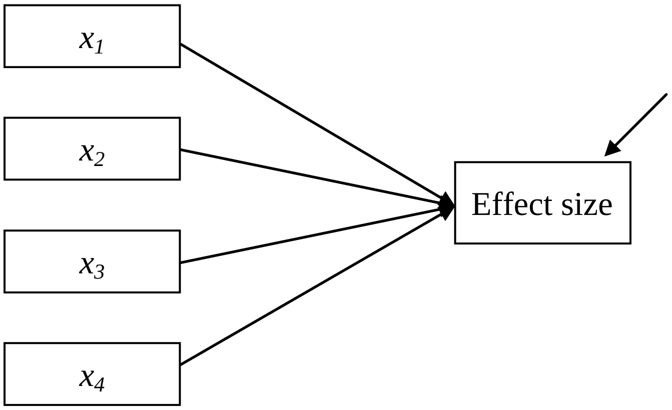
EPC = Evidence-based Practice Center; CBRG = Cochrane Back Review Group; RCT = randomized controlled trial

Analysis

Figure 2 depicts the basic hypothesis of the project: the assumption that there is an association between quality features of research studies and the size of the reported treatment effect. The arrows indicate the direction of effects. The figure also depicts the assumption that other variables apart from quality will affect effect sizes, as represented by the arrow on the right. These other variables include the true effect of the intervention as well as other potential influences; quality variables may explain part of the reported effect sizes, but there are other and possibly more important factors that are not quality related (e.g., the efficacy of the treatment). The analysis covers descriptive information on the datasets, an evaluation of the association between quality and effect sizes, and an analysis of potential moderators and confounders to investigate which factors influence the association between quality criteria and effect sizes.

The three datasets were often used to replicate results obtained in one dataset to test the robustness of effects across datasets; some analyses were only possible in one or two datasets. The initial intention to combine all three datasets to allow more powerful analyses was not feasible due to differences in outcome measures (all RCTs in dataset 3 used dichotomous outcomes, to transform all outcomes into continuous measures was considered problematic).

Figure 2. Quality indicators and effect sizes



Since this analysis plan involves multiple testing, we considered several methods for accounting for this; however these are not appropriate when tests are correlated. In addition, there is debate about the range to which multiple testing corrections should be employed (for an analysis, a study) and each of these would lead to different conclusions. All statistical multiple testing approaches lead to substantial loss of power (Bland and Altman, 1995). We therefore chose not to employ any of the methods to "correct" for multiple testings. Instead our results need to be interpreted with more caution, as a result of multiple testing.

Data Description

The three datasets were derived through different means and differed in a number of ways. First, we investigated if there were systematic differences related to the level of quality within the datasets. The level of quality may vary between clinical fields as the clinical areas may have different standards or awareness of quality features. The quality of published RCTs may have improved since the publication of the Consolidated Standards of Reporting Trials (CONSORT) statement in 1996 so another variable we explored further was the year of publication of studies included in each dataset.

To describe the internal consistency of the quality items, inter-item correlations and the Cronbach's alpha statistic for an overall quality scale were computed in each dataset. The Pearson correlations across items were inspected for consistency (are the individual quality features positively correlated) but also to detect high inter-item correlations (e.g. above 0.5) as an indicator for conceptual overlap (the answer in one item lets us predict the answer in another item).

All of the items score quality features. It is possible that the features are independent of each other (blinding of outcome assessors is not necessarily related to the similarity of the co-interventions). However, empirically the presence of one quality indicator might increase the likelihood that a second quality criterion is also fulfilled. For example, a study that used an appropriate method for a randomization sequence may also be likely to have employed an appropriate method to guarantee allocation concealment. Finally, theoretically, it is also possible that the individual items are indicators of an underlying factor representing "quality." A quality RCT is more likely to show several fulfilled quality criteria. Individual quality items may be indicators of this underlying quality factor.

We also used the individual quality items to create a sum scale. This overall quality score was computed by calculating the average quality scores across all items, with all items being weighted equally. Cronbach's alpha values range from 0 to 1; alpha coefficients above 0.7 indicate internal consistency. The Cronbach's alpha statistic was exploratory and was chosen as a measure with well-known properties, not because we assume a shared overarching latent quality factor. The included quality features may still be conceptually independent from another and may not represent items from the same item pool of a shared latent factor.

We also used factor analysis to explore the structure of the relationships between the 11 items. Conventional exploratory factor analysis attempts to find latent factors which explain the covariance between a set of items. Factor analysis assumes an underlying factor that is hypothesized to influence a number of observed variables, that is, the individual items. Factor analysis can show whether all included items can be explained through one underlying factor (e.g., "quality"), whether there are clusters of items representing different quality aspects, or whether all 11 items are unrelated and represent unique features. Conventional factor analysis only takes the pattern of quality scores across items into account; this approach does not incorporate the relationships with an outcome (such as effect size). We used an extension of factor analysis, a multiple indicator multiple cause (MIMIC) model, which allows us to model the relationships between the items in an exploratory fashion, and simultaneously examine the relationship between the latent variables that were identified and the outcome of interest (in this case, the effect size of the study). The factor analysis hence takes the inter-item relationships as well as the strength of association with effect sizes into account.

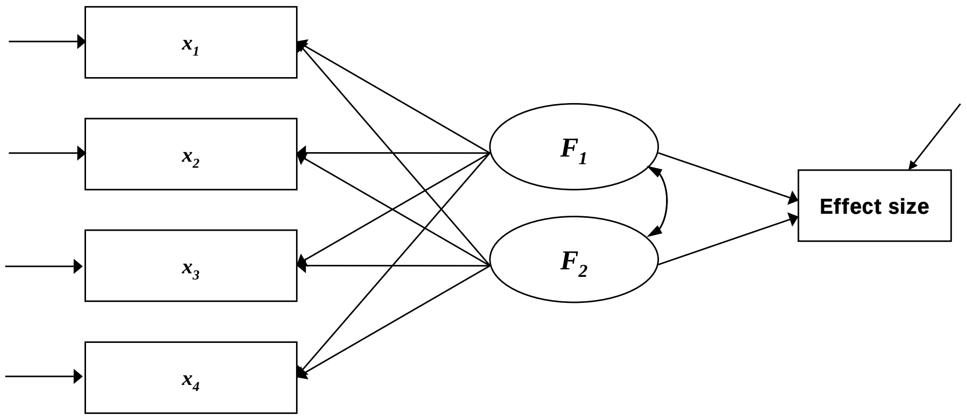
The path model shown in Figure 3 below is a simplified diagrammatic representation of the model assuming four indicators of quality (x_1 to x_4), that is, individual quality features. Single-headed arrows are regression paths—the four indicators of quality are hypothesized to be explained by two latent variables, F_1 and F_2. The two latent (unmeasured) variables represent distinct broad quality domains but are not necessarily completely independent from each other either; we assume in our model that they are correlated (the two-headed, curved arrow indicates this).

We hypothesize that the covariances between *x* variables are accounted for by the factors, the latent variables. We assume the latent (unmeasured) factors (F1, F2) are responsible for the majority of variation in individual quality criteria, and that these latent variables are also predictors of effect size. The indicator variables, that is, the individual quality items, are not conceptualized as being correlated; they can be independent from another, such as blinding and similarity of cointerventions. The partial correlation between individual quality indicators and the effect size is diminished to zero when controlling for the latent factors.

In summary, the effect size reported in the trials is regressed on the latent variables—thus quality is indicated by the x-variables (individual quality features), but the latent variables (unmeasured, broad quality domains) are hypothesized to predict variance in the effect size. It has to be kept in mind that variables other than quality will affect effect sizes, as represented by the arrow on the right.

To identify the appropriate number of latent factors that are required to account for the data, we employed fit indices (χ^2, comparative fit index, [CFI] and root mean square error of approximation [RMSEA]). We tested a series of models, each time increasing the number of factors and comparing the improvement of the model fit. This approach is used to determine the smallest number of factors that can account for the data.

Figure 3. Model assuming latent factor influencing effect size



The factor analysis solution is more parsimonious and enables a large number of items to be reduced to a smaller number of underlying factors. Factor analysis allows summarizing results across items without reducing the quality data to a summary score. However, the analysis should be considered descriptive as data are not weighted by standard error as is conventional in meta-analysis.

Association Between Quality and Effect Sizes

We investigated the association between quality and effect sizes in a number of ways. First, the differences between results in studies that met a quality criterion were calculated for each of the 11 quality features. Secondly, a summary score was calculated across all quality components and a linear relationship between quality and effect sizes was investigated. Third, the associations based on empirically derived factor scores was tested, the factor structure took the inter-correlations between items and their effects on outcomes into account. Fourth, we explored different cutoffs of quality scores according to the number of quality components met.

For all analyses we differentiated quality items scored “yes” and those with the quality item scored “not yes” which included the answer “no” and “unclear” unless otherwise stated.

Individual Criteria

In the first two datasets (back pain, EPC reports) we compared the effect sizes of studies with the quality item scored “yes” and those with the quality item scored “not yes” for each of the 11 quality features. The difference in effect sizes between these two subgroups per feature was used as a measure of bias. The difference was estimated using meta-regression (Berkey et al., 1995). A meta-regression was conducted separately for each quality feature. The coefficient from each regression estimates the difference in effect sizes between those studies with the quality feature scored “yes” versus “not yes.” A difference with a significance level of p<0.05 was considered statistically significant.

In the third dataset, the published “pro-bias” dataset, all studies used dichotomous outcomes. An odds ratio below 1 indicates the treatment group is doing better than the control. For the analysis we compared odds ratios (ORs) of studies where the quality criterion was either met or not met and computed the ratio of the odds ratios (ROR). The ROR is OR_{no}/OR_{yes} where OR_{no} is the pooled estimate of studies without the quality feature and OR_{yes} is the pooled estimate of studies where the quality criterion is met.

Note that the interpretation of reported differences of the first two datasets differs from that of the third one. In the first two datasets a negative difference coefficient indicated that studies with the quality item scored “yes” have smaller effect sizes than those that scored “not yes.” Hence, a negative difference indicates that the higher quality RCTs report less pronounced treatment effects. In the third dataset a ROR of being less than 1 indicates that high quality studies reported smaller treatment effects (i.e., the OR closer to 1) than low quality studies.

We compared results based on a random effects meta-regression, and a fixed effects model in order to be able to match results reported in the literature.

Sum Scores

The sum of the quality items scored “yes” was calculated across all 11 items with all items contributing equally to the total score. To assess a linear relationship between overall quality and effect size, reported outcome results were regressed on the sum score. A simple linear relationship indicates that the reported treatment effects increase the lower the quality level. A level of $p<0.05$ was considered statistically significant.

Factor Scores

We used the empirically derived factor scores representing broad quality domains and regressed effect sizes on these factors. The factor scores were based on the inter-item relationships as well as their effects, that is, the association with the study results that provides a description of distinct groups of items. The analysis was equivalent to the sum score analysis.

Cutoffs

Different cutoffs depending on the number of criteria met were explored to differentiate high and low quality studies. The difference in effect sizes of studies above and below possible thresholds was investigated. The statistical analysis followed the approach outlined under (1).

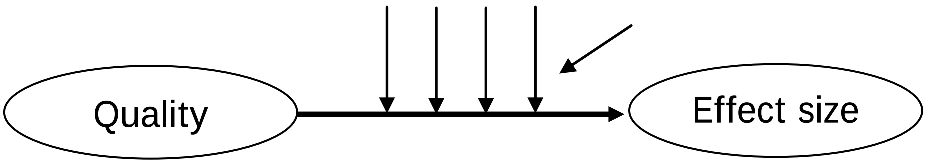
The different methods of establishing associations between quality and effect sizes were exploratory and we did not a priori assume consistent results across methods. For example, a simple linear relationship between a total quality scale and effect sizes will not necessarily be present even when individual quality features show large associations with effect sizes; the internal consistency across items was one of the issues under investigation.

The analyses were conducted separately in each of the three datasets. Each dataset consisted of trials included in up to 12 meta-analyses. We did not correct for clustering in analyses within datasets. We do not assume nonindependence of RCTs within meta-analyses since the selection into the meta-analysis happened after the event (when the study was already conducted and published).

Moderators and Confounders

Effect sizes are influenced by many variables, not just the methodological quality of the research study. In addition, we have to assume from conflicting literature results that there are factors that influence the relationship between methodological quality and the effect size. Figure 4 shows a model that assumes factors influencing the association between quality and effect sizes and indicates that effect sizes are also influenced by other variables independent from trial quality.

Figure 4. Moderators and confounders



Two effects need to be considered: confounding effects and moderating effects. These are of particular relevance in dataset 2, where papers are selected from a wide range of clinical topics and interventions.

Confounding effects occur when the quality of trials is not equally distributed across areas of study, resulting in a correlation between quality and area of study. This correlation can lead to erroneous conclusions if the area of study is not incorporated as a covariate. In extreme cases, this correlation can lead to counter-intuitive results, an effect known as Simpson's paradox. The example in Table 1 considers two areas of study, labeled A and B, and a measure of quality, such as randomization, which is either achieved or not achieved, giving four combinations. The effect sizes are given in the table below. Within study area A, the effect size is 0.1 higher when the quality measure is not achieved. Similarly, within study area B the effect size is 0.1 higher when the quality rating is not achieved. However, studies in Area B have higher effect sizes on average (0.25) than studies in Area A (0.15), and studies in Area B are much more likely to have achieved the quality rating. This confounding means for subpopulations of studies the result is in one direction, but for the whole population the result is in the other direction.

Table 1. Example confounding effect showing Simpson's paradox

		Substantive Area A	**Substantive Area B**	**Mean (Weighted) Effect Size**
Quality Achieved	Yes	Effect: 0.1 N: 2	Effect: 0.3 N: 20	0.28
	No	Effect: 0.2 N: 20	N: 0.4 N: 2	0.22

The second potential issue is one of moderation. In the case of moderation, the causal effect for a quality rating varies between different substantive areas. We illustrate a moderator effect in Table 2. This example shows that for substantive area A, quality does not influence the effect size; however for area B there is a substantial influence of quality on effect size. To take the average quality association would be inappropriate when the influence differs across substantive areas (and would therefore be influenced by the number of studies identified in each area).

Table 2. Example moderator effect

		Substantive Area A	**Substantive Area B**
Quality Achieved	Yes	Effect: 0.1	Effect: 0.1
	No	Effect: 0.1	Effect: 0.4

The literature reports some conflicting results regarding the strength of association between quality features and effect sizes indicating that we have to assume factors that influence the relationship between the two variables. In this project we set out to investigate the influence of four variables: the size of the treatment effect, the condition that is being treated, the type of analyzed outcome and the variance in effect sizes across studies for the quality feature in question.

Variable 1: Size of Treatment Effect

We tested the hypothesis that the association of quality features and reported effect sizes varies according to the size of the overall treatment effect. Strong treatment effects may mask any effects of quality features on the individual study outcome. Similarly, an ineffective treatment may likewise yield the same result regardless of study quality. We computed the mean effect size for each included meta-analysis and added this variable to the regression models and compared results between two datasets.

Variable 2: Condition Being Treated

We tested the hypothesis that the association of quality features and effect size varies by condition. Under this hypothesis the selection of clinical conditions in a dataset determines whether or not an association between quality and effect size can, or cannot be shown. The underlying factors for this differential effect may remain unknown; we are only testing whether the association with quality features can be documented in one clinical area or groups of clinical areas but not in others.

The analysis was restricted to the large and diverse EPC report dataset (dataset 2, 165 trials). The back pain studies addressed a homogeneous condition. The third dataset was too small to investigate the effects for each of the 11 included conditions (most comparisons would be incomputable) and too unbalanced (only 2 out of 11 studies were not drug studies, only 1 meta-analysis was in pregnancy and childbirth).

Variable 3: Type of Outcome

We tested the hypothesis that the association of quality and effect sizes varies by type of outcome. Whether an association of quality and effect sizes can be shown may depend primarily on the investigated outcome. Some types of outcomes may be more susceptible to bias than others. More objective versus more subjective outcome measures may represent different kind of outcome types. Hypothesis 3 tests whether the association of quality features and effect size may vary by the type of analyzed outcome.

In the back pain dataset, the measured outcomes were all either subjective outcomes such as pain or outcomes involving clinical judgment such as “improvement,” so this set could not contribute to this moderator analysis. The outcomes in the EPC report dataset were more diverse. We distinguished automated data (hemoglobin A1c, high-density lipoprotein, and total cholesterol) versus other endpoints (Alzheimer’s Disease Assessment Scale cognition score, arthritis responders, reduction in seizures, pain, OCD improvement, weight loss, and depression scores). In the third dataset, we distinguished objective data such as death, pregnancy, and biochemical indicators of smoking cessation, from other endpoints of a more subjective nature or involving clinical judgment (response in ulcer healing or pain relief, bleeding complications, schizophrenic relapse, ulcer healing rate, affective relapse, and maintenance of ulcerative colitis remission).

Variable 4: Variance in Effect Sizes

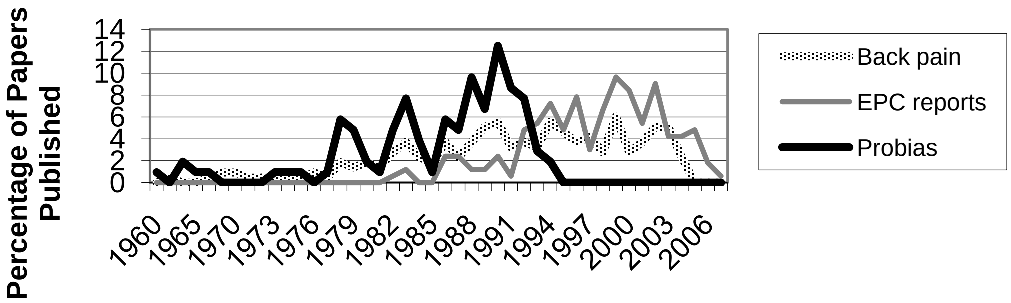
We tested the hypothesis that the association of quality features and effect sizes may depend on the variance in effect sizes across studies in a given dataset. In a dataset where there is a wide range of reported effect sizes across studies, quality may be more likely to explain differences in effect sizes across studies.

Results

Data Description

The years of publication of the included papers are shown in Figure 5.

Figure 5. Year of publication of included studies



EPC = Evidence-based Practice Center

The three datasets showed some differences: dataset 3 (published “pro-bias” dataset) included many older papers with a peak in the 1990s compared to the other datasets and all studies were published before 1996. The dataset 1 (back pain data) included mainly newer publications, several published in the last decade. The studies included in the Evidence-based Practice Center (EPC) reports were published over a large period of time, with no particular peak.

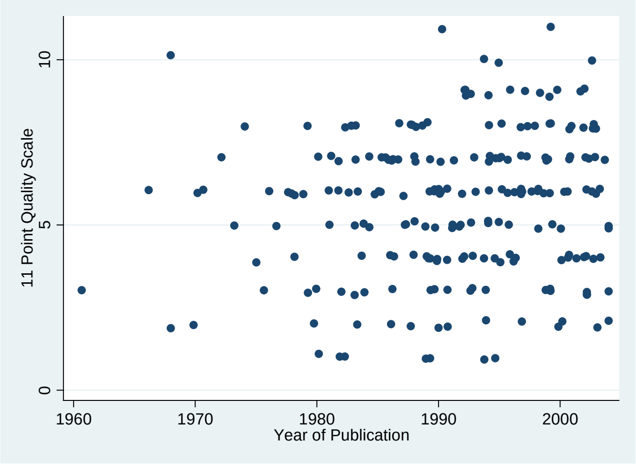
Relationship Between Total Quality Scores and Year of Publication

We investigated in each dataset the relationship between a quality sum score based on the mean of the assessed quality features and the year of publication. Figure 6 plots both variables.

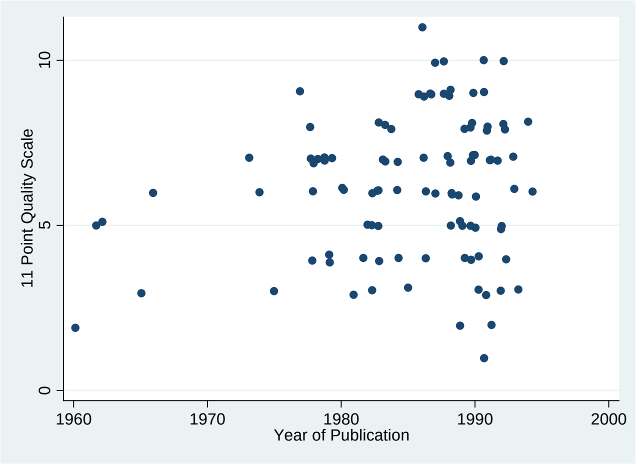
In addition, the difference in quality between pre- and post-Consolidated Standards of Reporting Trials (CONSORT) publications was tested (1996 used as cutoff). In the back pain dataset, the difference in total scores between pre- and post-CONSORT published trials was 0.58 (SE 0.32, p=0.07) on the 11-item scale. The quality of studies published after the introduction of the CONSORT statement was better but not statistically significantly higher. In the EPC report dataset the difference between pre- and post-CONSORT quality ratings was 1.35 (SE 0.32, p<0.001). All studies included in the third dataset were published before the introduction of CONSORT.

To ensure that the effect was not an artifact of the fact that quality of studies was improving over time anyway, regardless of CONSORT, we estimated the effect of time for papers published both pre- and post-CONSORT. These effects were not statistically significant.

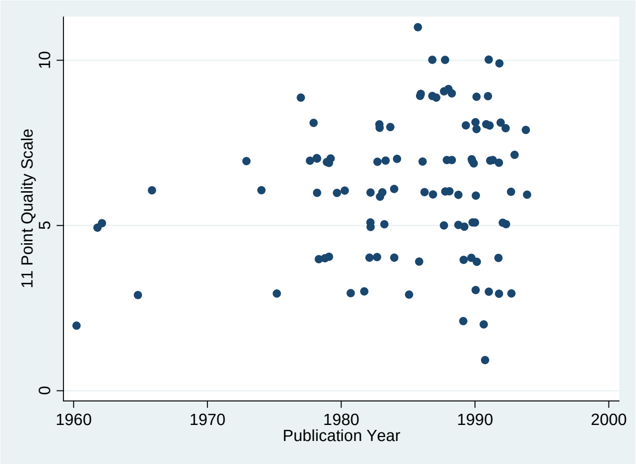
Figure 6. Total quality and year of publication



Dataset 1: Back pain



Dataset 2: EPC reports



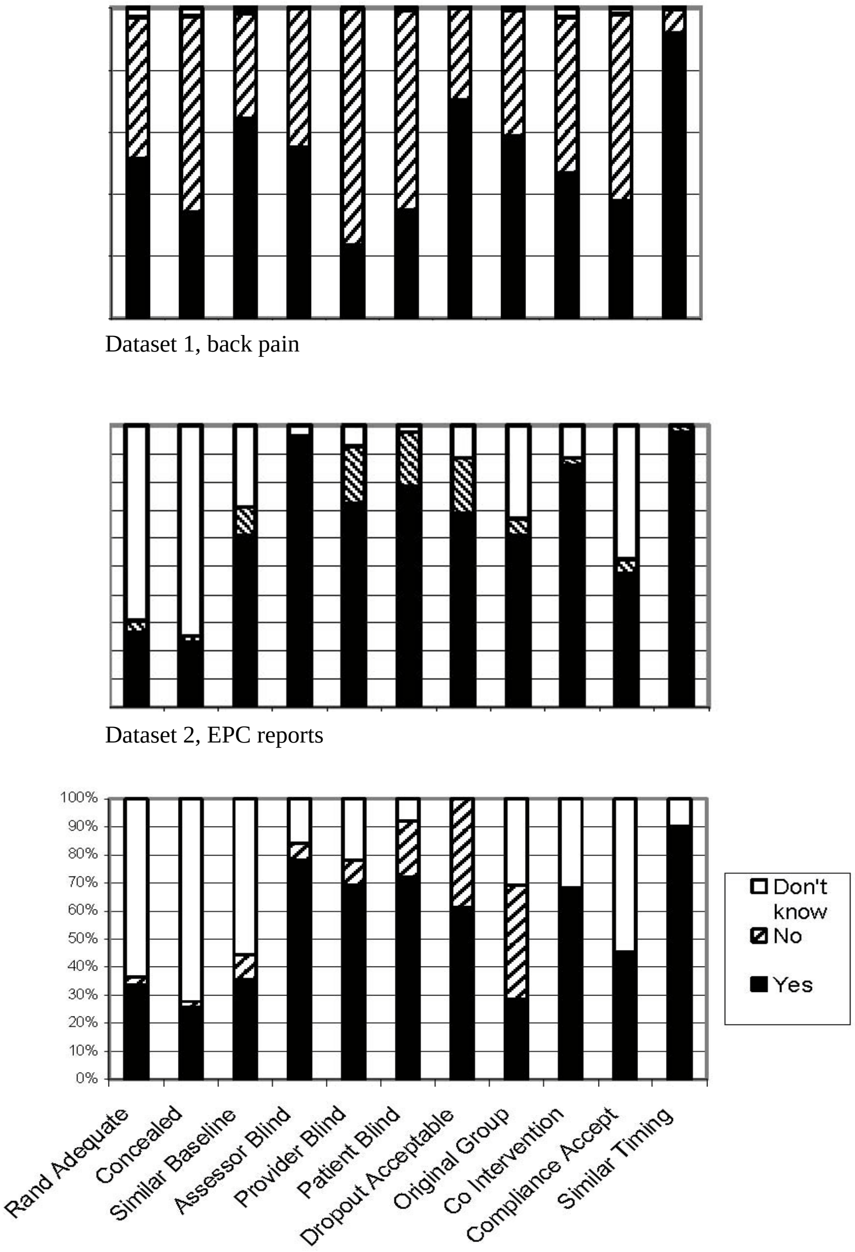
Dataset 3: “pro-bias”

Note: data points have been “jittered” to avoid overlap.

Quality of the Reporting

Figure 7 shows the distribution of answers to the quality items (yes, unclear, no). A “yes” is an indicator of high quality for all items (randomization sequence, allocation concealment, baseline similarity, outcome assessor blinding, care provider blinding, patient blinding, dropout rate and description, analysis in original group, intention to treat [ITT], cointerventions, compliance, and assessment timing), for example, the outcome assessors were blind.

Figure 7. Quality item answer distribution



Dataset 1, back pain

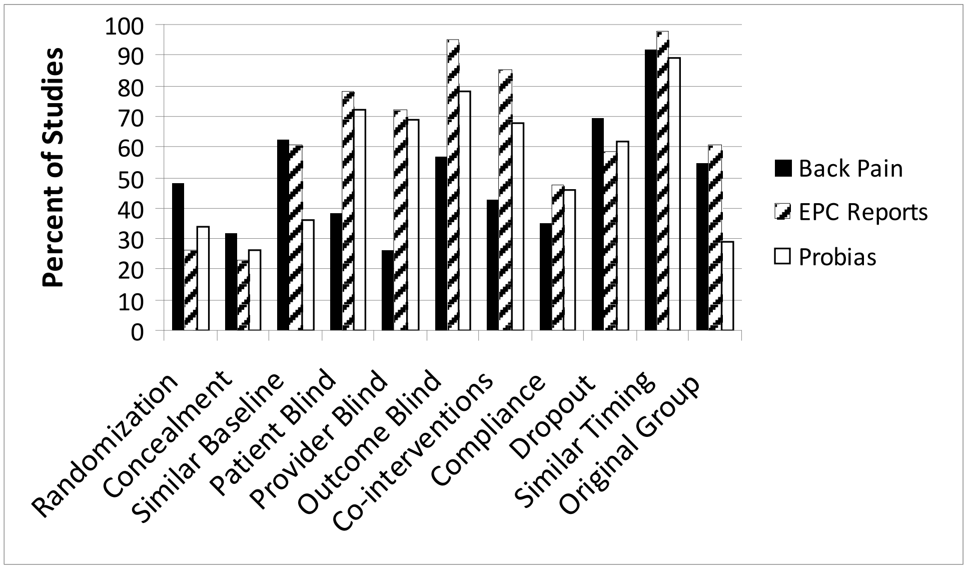
Dataset 2, EPC reports

Dataset 3, “pro-bias”

In the back pain dataset, the presence or absence of quality features is relatively evenly distributed for most items. Patient and provider blinding was not very common in the included trials and presumably often impossible due to the nature of the interventions. Similar timing of outcome measure assessment in the treatment and the control group was common, but there were a few deviations. The studies included in the EPC reports showed less variation across items. Many quality features were either present or there was not enough information to judge the individual quality feature. The answer “no” was very rare. The “unclear” answer was very common in dataset 2 (EPC reports) and 3 (published “pro-bias” dataset) indicating that the original studies did not report enough information to judge the quality feature. Very few trials scored negatively for the assessed quality features, that is, the reviewer had enough information to know that the design feature was not adhered to (e.g., the patient was not blinded). In datasets 2 and 3 there was virtually no variance in the item “Was the timing of the outcome assessment similar in all groups?” across studies, indicating that this quality feature may be unique to back pain trials.

Figure 8 allows a comparison of “yes” answers across the three datasets.

Figure 8. Criterion met across datasets



EPC = Evidence-based Practice Center

The level of criteria met was highest in the EPC report dataset for the blinding items, similarities of cointerventions, similar timing, and the analysis in the original group assignment (ITT analysis). The lowest quality level across quality criteria was generally observed in the third dataset, which contains older studies, all published before the CONSORT statement.

Intercorrelations Quality Features

Although conceptually presumably independent, in practice studies that pay attention to one quality feature (e.g., allocation concealment) often do so also for others (e.g., using an adequate method of generating a randomization sequence). To trace the empirical interrelatedness of the quality items, Tables 3–5 show the inter-item correlations of quality features in each of the three datasets.

Table 3. Dataset 2 Inter-item correlations dataset 1 (back pain)

Randomization Adequate	1.00									
Allocation Concealment	0.53	1.00								
Similar Baseline	0.19	0.24	1.00							
Assessor Blind	0.01	0.09	0.00	1.00						
Care Provider Blind	-0.30	-0.21	-0.08	0.46	1.00					
Patient Blind	-0.20	-0.17	-0.08	0.55	0.68	1.00				
Acceptable Dropout Rate	0.10	0.07	0.15	0.05	0.08	-0.10	1.00			
Original Group (ITT)	0.23	0.27	0.12	-0.02	-0.05	-0.11	0.40	1.00		
Similar Cointerventions	0.20	0.19	0.13	0.14	-0.01	-0.04	0.14	0.13	1.00	
Acceptable Compliance	0.28	0.31	0.25	-0.03	-0.18	-0.14	0.07	0.20	0.31	1.00
Similar Timing	-0.04	0.10	0.15	-0.06	0.03	-0.01	0.13	0.13	0.12	0.12
	Rand Adequate	**Concealed**	**Similar Baseline**	**Assessor Blind**	**Provider Blind**	**Patient Blind**	**Dropout Acceptable**	**Original Group**	**Co-Intervention**	**Compliance Accept**

ITT = intention to treat

The mean inter-item correlation in this dataset was 0.1. Some of the items were substantially inter-correlated, for example, was the treatment allocation concealed correlated highly with an adequate randomization sequence, and if studies reported patient blinding, the studies tended to also report provider and outcome assessor blinding. The majority of features showed coherence but did not indicate that items were redundant and the information for one item was contained in another. There were a few negative correlations; the only noteworthy correlation was that the studies that reported an adequate randomization procedure stated that the care providers were not blinded (often impossible in this dataset given the interventions).

Table 4. Dataset 2 Inter-item correlations dataset 2 (EPC reports)

Randomization Adequate	1.00									
Allocation Concealment	0.74	1.00								
Similar Baseline	0.03	0.03	1.00							
Assessor Blind	0.01	0.06	-0.07	1.00						
Care Provider Blind	0.13	0.15	-0.17	0.36	1.00					
Patient Blind	0.05	0.11	-0.15	0.15	0.78	1.00				
Acceptable Dropout Rate	0.12	0.08	-0.08	-0.02	0.05	0.09	1.00			
Original Group (ITT)	0.15	0.23	-0.04	-0.01	0.41	0.36	0.00	1.00		
Similar Cointerventions	0.09	0.02	0.09	-0.01	0.17	0.20	-0.07	0.13	1.00	
Acceptable Compliance	0.06	0.06	0.00	-0.01	0.08	0.09	0.19	0.07	0.08	1.00
Similar Timing	0.01	-0.01	0.12	0.15	0.17	0.20	0.03	0.03	0.38	0.15
	Rand Adequate	**Concealed**	**Similar Baseline**	**Assessor Blind**	**Provider Blind**	**Patient Blind**	**Dropout Acceptable**	**Original Group**	**Co-Intervention**	**Compliance Accept**

ITT = intention to treat

The mean inter-item correlation in the EPC reports was r = 0.11. Adequate sequence of randomization and concealment of treatment allocation were highly intercorrelated, as were provider and patient blinding.

Table 5. Dataset 3 Inter-item correlations dataset 3 ("pro-bias")

Randomization Adequate	1.00									
Allocation Concealment	0.49	1.00								
Similar Baseline	0.34	0.13	1.00							
Assessor Blind	0.02	0.15	0.25	1.00						
Care Provider Blind	0.07	0.00	0.28	0.69	1.00					
Patient Blind	-0.02	-0.04	0.24	0.53	0.79	1.00				
Acceptable Dropout Rate	0.13	0.04	0.16	-0.12	-0.03	-0.08	1.00			
Original Group (ITT)	0.01	-0.03	0.16	-0.14	-0.05	-0.04	0.00	1.00		
Similar Cointerventions	0.09	-0.08	0.34	0.10	0.19	0.15	-0.01	0.06	1.00	
Acceptable Compliance	-0.03	0.00	-0.07	-0.14	-0.12	-0.14	0.06	0.03	0.07	1.00
Similar Timing	0.05	0.06	0.26	0.12	0.32	0.42	0.05	0.01	0.17	0.07
	Rand Adequate	**Concealed**	**Similar Baseline**	**Assessor Blind**	**Provider Blind**	**Patient Blind**	**Dropout Acceptable**	**Original Group**	**Co-Intervention**	**Compliance Accept**

ITT = intention to treat

The mean inter-item correlation in the third dataset was 0.11, and the pattern was very similar to the two other datasets.

In addition to the newly proposed quality features, we also applied the Jadad scale and the quality features suggested by Schulz et al. (1995) including allocation concealment. Table 6 shows the intercorrelations between the newly proposed items and the features scored according to the Jadad scale instructions and Schulz's original instructions. Correlations in bold indicate corresponding quality domains. As expected, there were strong correlations between Cochrane Back Review Group (CBRG) Internal Validity items and corresponding items for Jadad and Schulz—items describing randomization, concealment, blinding, and dropouts.

Table 6. Correlation of criteria with Jadad and measures proposed by Schulz

Quality Feature	Jadad Randomization	Jadad Blinding	Jadad Withdrawals	Jadad Total	Schulz Concealment	Schulz Sequence	Schulz Analysis	Schulz Blinding
Was the method of randomization adequate?	**0.86**	0.13	0.04	0.50	0.39	**0.91**	0.00	0.07
Was the treatment allocation concealed?	0.44	0.12	-0.01	0.29	**0.84**	0.46	0.06	0.04
Were the groups similar at baseline regarding the most important prognostic indicators?	0.31	0.33	0.16	0.43	0.08	0.28	0.05	0.32
Was the outcome assessor blinded?	0.13	**0.65**	-0.09	0.47	0.09	0.03	-0.20	**0.74**
Was the care provider blinded?	0.13	**0.82**	0.05	0.64	-0.05	0.06	0.03	**0.93**
Were patients blinded?	-0.02	**0.76**	0.04	0.53	-0.09	-0.03	0.11	**0.87**
Was the dropout rate acceptable?	0.12	0.02	**0.76**	0.34	0.04	0.15	-0.27	-0.04
Were all randomized participants analyzed in the group to which they were originally assigned?	-0.03	-0.12	0.13	-0.05	0.02	-0.03	**0.63**	-0.05
Were co-interventions avoided or similar?	0.14	0.22	0.08	0.25	-0.18	0.07	-0.04	0.23
Was the compliance acceptable in all groups?	0.02	0.00	0.18	0.08	0.09	-0.04	0.13	-0.10
Was the timing of the outcome assessment similar in all groups?	0.11	0.31	0.08	0.30	0.06	0.02	0.10	0.35

Note: in bold are correlations reflecting similar constructs.

The correspondence was not perfect, since although assessing similar quality domains, the scoring rules are not identical across quality scoring systems. For example, the CBRG provides a number of rules for when blinding can be assumed when blinding was not explicitly reported in the study (e.g., when automated data are concerned). Schulz's item "Inclusion in the Analysis of All Randomized Participants" instructs that the item should be answered in the affirmative when the publication reports "or gives the impression" that no exclusions have taken place ("often not explicit"), whereas the corresponding CBRG item requires an explicit statement in the text or explicit data; otherwise the item will be scored "unclear."

Internal Consistency

In the back pain dataset, the Cronbach's alpha coefficient of a summary scale was 0.56, in the EPC reports alpha was 0.55, and in the third dataset alpha was 0.61, indicating in all three datasets only moderate internal consistency. The level of consistency does not indicate that all items are measuring the same construct.

To investigate whether the internal consistency was mainly influenced by one or two outlying items, the alpha coefficient was computed excluding each item in turn (alpha if item deleted analysis), depicted in Table 7. This analysis can also show whether there are items that do not add any information that is already captured though other items (in that case, the scale alpha does not drop although the item is removed from the scale).

Table 7. Alpha if item deleted

Quality Feature	Scale α if Item Deleted Back Pain α=0.56	Scale α if Item Deleted EPC Reports α=0.55	Scale α if Item Deleted Published Dataset "Pro-bias" α=0.61
Was the method of randomization adequate?	0.53	0.49	0.54
Was the treatment allocation concealed?	0.50	0.49	0.56
Were the groups similar at baseline regarding the most important prognostic indicators?	0.53	0.61	0.48
Was the outcome assessor blinded?	0.52	0.55	0.53
Was the care provider blinded to the intervention?	0.56	0.46	0.49
Were patients blinded?	0.57	0.47	0.51
Was the dropout rate described and acceptable?	0.53	0.57	0.59
Were all randomized participants analyzed in the group to which they were originally assigned?	0.52	0.50	0.60
Were cointerventions avoided or similar?	0.51	0.53	0.54
Was the compliance acceptable in all groups?	0.52	0.55	0.61
Was the timing of the outcome assessment in all groups similar?	0.55	0.54	0.52

EPC = Evidence-based Practice Center

The "alpha if item removed" scores indicated in the back pain dataset that one of the blinding items may be unnecessary since its absence does not decrease alpha. All other items did not affect the total score substantially, there was also no indication that one particular item was the "odd one out," not related to an overall quality score constituted by these 11 quality features. In the EPC projects, removing the items that concern the similarity of the baseline would raise alpha slightly, and the outcome assessor blinding item does not add any information that is not already captured by (presumably) the other blinding items. The removal of the compliance item would not lower the Cronbach's alpha value in the third dataset, indicating that this item does not contribute to increased reliability of a total scale.

Factor Analysis

To investigate the structure of relationships between quality features we fitted one, two, and three factor models in a multiple indicator multiple cause (MIMIC) model to each dataset. The factors group related items, in terms of intercorrelations as well as in their strength of association with effect sizes. For each dataset, we estimated three models—a single quality factor model, two quality factors, and three quality factors. (Figure 3 shows a two quality factor model). Each model was assessed using a range of fit measures, which indicate the degree of misfit between the model and the data. The χ^2 test should be nonsignificant in a well-fitting model, the comparative fit index should be over 0.95, and the root mean square error of approximation should be less than 0.05.

In meta-analysis, the studies are weighted by the standard error of the estimate. In fitting a MIMIC model, techniques have not been developed that allow us to estimate weights for studies separately; hence, the relationships between effect size and quality factors should be interpreted in this light.

The fit indices for the datasets are shown in Table 8. In the back pain dataset, the patient blinding item and similar timing had to be removed due to collinearity, in the EPC report set the assessor blinding and similar timing had to be removed, and in the third dataset the assessor blinding item was dropped due to lack of variance. A model assuming three factors gave a good fit to the data in all three datasets.

Table 8. Fit indices

Fit Indices	Dataset 1: Back Pain			Dataset 2: EPC Reports			Dataset 3: "Pro-bias"		
	1 Factor	2 Factor	3 Factor	1 Factor	2 Factor	3 Factor	1 Factor	2 Factor	3 Factor
χ^2 (df) p	169 (23) <0.001	N/C	19.1 (13) 0.120	71 (15) <0.001	N/C	11.6 (12) 0.47	52 (23) <0.001	27 (21) 0.18	16.1 (19) 0.72
CFI	0.59		0.98	0.96		1.00	0.71	0.94	1.000
RMSEA	0.171		0.47	0.15		0.000	0.11	0.053	0.000

df = degrees of freedom; N/C = model failed to converge; CFI = comparative fit index; RMSEA = root mean square error of approximation

The factor loadings of the individual quality features on the latent factors are shown in Table 9. The largest loading of each item in each dataset is highlighted in bold. Factor loadings are the correlations of each quality feature with the factor. Factor loadings constitute the factors.

Table 9. Standardized factor loadings

Quality Feature	Back Pain			EPC Reports			Published Set "Pro-bias"		
	F1	F2	F3	F1	F2	F3	F1	F2	F3
Randomization adequate	**0.82**	-0.07	0.02	**0.94**	-0.04	0.06	**0.77**	0.45	-0.01
Allocation concealment	**0.92**	0.08	0.02	**0.99**	0.04	-0.02	**1.05**	0.00	0.30
Similar baseline	**0.41**	0.00	0.18	0.17	**-0.28**	-0.16	0.35	**1.06**	0.01
Assessor blind	0.22	**1.01**	-0.08				0.18	-0.01	**0.95**
Care provider blind	-0.42	**0.83**	0.14	0.03	**0.98**	-0.01	0.00	0.12	**0.96**
Patient blind				-0.07	**0.99**	0.06	-0.16	0.15	**0.83**
Acceptable dropout rate	0.00	0.00	**0.99**	0.03	0.01	**0.95**	0.16	**0.29**	-0.24
Original group (ITT)	0.31	-0.08	**0.58**	0.27	**0.56**	-0.15	-0.03	**0.33**	-0.30
Similar co-interventions	**0.43**	0.20	0.17	0.11	**0.32**	-0.18	-0.07	**0.61**	-0.05
Acceptable compliance	**0.59**	-0.05	0.13	0.05	0.12	**0.25**	-0.01	0.06	**-0.25**
Similar timing							0.00	**0.42**	0.40

Note: loadings in bold are significant (p<0.05).
EPC = Evidence-based Practice Center; F = factor; ITT = intention to treat

Factor one in the back pain dataset consisted mainly of the randomization sequence item, allocation concealment, and the compliance item. Baseline similarity and the similarity of co-interventions also loaded on this item. Factor two represented blinding. Factor three was made up of the acceptable dropout item across datasets and the ITT item (original group) in two out of the three datasets. The correlations between factors were not statistically significant (F1, F2: -0.14 [p=.224]; F1, F3: 0.13 [p=0.281]; F2, F3: 0.11 [p=0.252]).

In the EPC report, dataset factor one consisted of randomization sequence and allocation concealment. Items relating to blinding loaded on factor 2 as did the original group, co-intervention and similar baseline items. The only item with a high loading on factor 3 was the dropout measure. The correlations between factors were not statistically significant (F1, F2: 0.24 [p=.056]; F1, F3: 0.15 [p=.230]; F2, F3: 0.09 [p=.446]).

In the third dataset, the randomization item and allocation concealment loaded again on factor 1. Factor 3 was the blinding factor, but factor 2 consisted mainly of similar baseline and similar cointervention reporting and a couple of other items also loaded on this factor. The correlations between factors were not statistically significant (F1, F2: -0.22 [p=.553]; F1, F3: -0.20 [p=.392]; F2, F3: -0.42 [p=0.019]).

Similarities across all three datasets were that the randomization and concealment items formed one "treatment allocation" factor. Blinding constituted another robust factor across datasets, independent from the treatment allocation factor. In each dataset a third factor had to be assumed accounting for additional variance not covered by the two other factors.

Association Between Quality and Effect Sizes

Dataset 1: Back Pain Trials

As reported previously (van Tulder et al., 2009), studies included in the CBRG that scored positive for a quality item reported smaller effect sizes compared with trials that did not fulfill the criterion. The differences were not statistically significant but the included features showed consistency across domains with 10 out of 11 features indicating this effect, as depicted in Table 10.

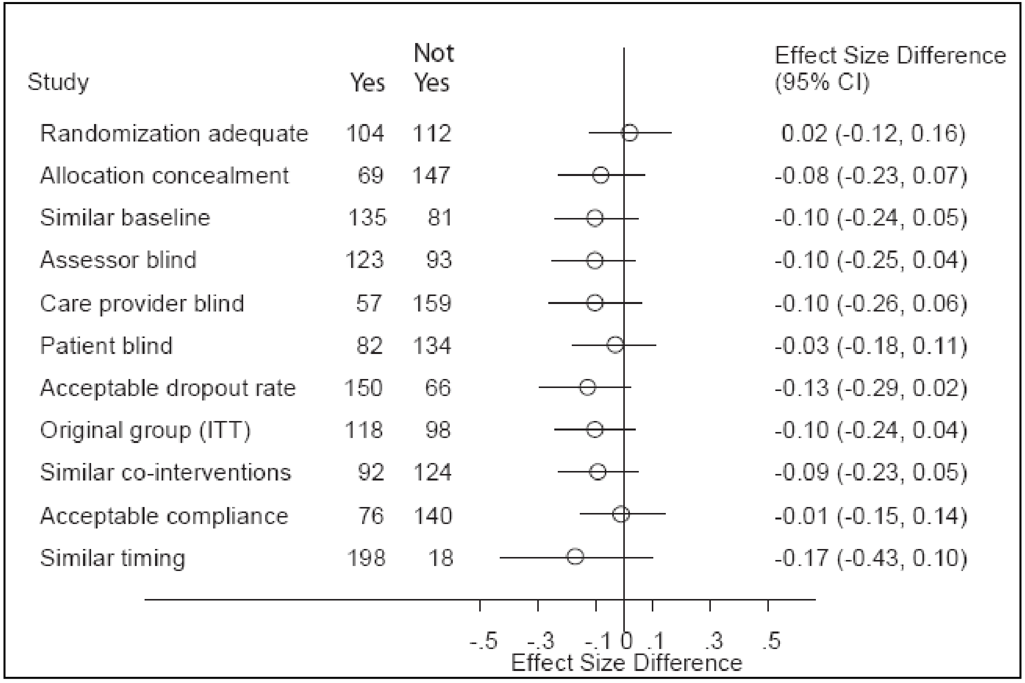
Table 10. Difference in effect sizes dataset 1 (back pain)

Quality Feature	Number Criterion Met	Number Criterion Not Met	Effect Size in Trials With Criterion Met		Effect Size in Trials With Criterion Not Met		Effect Size Difference	
			ES	95% CI	ES	95% CI	ESdiff	95% CI
Randomization adequate	104	112	0.51	(0.41, 0.61)	0.49	(0.40, 0.59)	0.02	(-0.12, 0.16)
Allocation concealment	69	147	0.45	(0.33, 0.57)	0.53	(0.45, 0.62)	-0.08	(-0.23, 0.07)
Similar baseline	135	81	0.47	(0.38, 0.55)	0.57	(0.45, 0.68)	-0.10	(-0.24, 0.05)
Assessor blind	123	93	0.46	(0.37, 0.55)	0.56	(0.46, 0.67)	-0.10	(-0.25, 0.04)
Care provider blind	57	159	0.43	(0.30, 0.56)	0.53	(0.45, 0.61)	-0.10	(-0.26, 0.06)
Patient blind	82	134	0.48	(0.37, 0.60)	0.52	(0.43, 0.60)	-0.03	(-0.18, 0.11)
Acceptable dropout rate	150	66	0.46	(0.38, 0.55)	0.60	(0.47, 0.73)	-0.13	(-0.29, 0.02)
Original group (ITT)	118	98	0.46	(0.37, 0.55)	0.56	(0.45, 0.67)	-0.10	(-0.24, 0.04)
Similar cointerventions	92	124	0.45	(0.35, 0.56)	0.54	(0.45, 0.63)	-0.09	(-0.23, 0.05)
Acceptable compliance	76	140	0.50	(0.39, 0.61)	0.51	(0.42, 0.59)	-0.01	(-0.15, 0.14)
Similar timing	198	18	0.49	(0.42, 0.56)	0.66	(0.40, 0.92)	-0.17	(-0.43, 0.10)

ES = effect size; CI = confidence interval; ESdiff = effect size difference; ITT = intention to treat

Figure 9 depicts the difference in effect sizes between studies meeting the individual quality criterion and those that do not. A negative effect size difference indicates possible bias; high-quality studies (those that meet the quality criterion) reported smaller effect sizes.

Figure 9. Difference in effect sizes based on quality features dataset 1 (back pain)



Study	Yes	Not Yes	Effect Size Difference (95% CI)
Randomization adequate	104	112	0.02 (-0.12, 0.16)
Allocation concealment	69	147	-0.08 (-0.23, 0.07)
Similar baseline	135	81	-0.10 (-0.24, 0.05)
Assessor blind	123	93	-0.10 (-0.25, 0.04)
Care provider blind	57	159	-0.10 (-0.26, 0.06)
Patient blind	82	134	-0.03 (-0.18, 0.11)
Acceptable dropout rate	150	66	-0.13 (-0.29, 0.02)
Original group (ITT)	118	98	-0.10 (-0.24, 0.04)
Similar co-interventions	92	124	-0.09 (-0.23, 0.05)
Acceptable compliance	76	140	-0.01 (-0.15, 0.14)
Similar timing	198	18	-0.17 (-0.43, 0.10)

CI = confidence interval; ITT = intention to treat

In terms of an established quality measure, differences in effect sizes between low and high quality studies were -0.14 (p<0.05, random effects meta-regression) for items constituting the Jadad scale. High-quality studies reported a mean estimated effect size of 0.45, while low-quality studies reported a mean estimated effect size of 0.60. The fixed-effects model effect size difference was -0.09 (p<0.05).

Summary and Factor Scores

To explore a linear effect of quality on effect size, we regressed the effect sizes on a total quality score value which we had computed for each study. The total quality score was obtained by equally weighing each of the 11 quality components to contribute to an overall quality score. The linear effect of the total quality scores across studies was negligible and not statistically significant (-0.04, SE 0.018, p=0.053, 95% CI: -0.073, 0.005).

When effect size was regressed on each of the three factors established in the factor analysis (each factor representing empirical clusters of quality features), the results were also not statistically significant. Standardized effects were for factor 1: 0.07 (p=0.699), factor 2: -0.23 (p=0.077) and factor 3: -0.15 (p=0.245).

Statistically significant results of quality were shown when applying a cutoff of 5 or 6 quality items fulfilled, the difference in effect size between low and high quality studies was 0.20 (van Tulder et. al., 2009). For this analysis, we used a total quality score per study and applied a cut-off empirically distinguishing high- and low-quality studies, depicted in Table 11.

Table 11. Comparison of different quality cutoffs using a total score dataset 1 (back pain)

Cutoff	Number Equal or Above Cutoff	Number Below Cutoff	High Quality		Low Quality		Difference	
			ES	95% CI	ES	95% CI	ESdiff	95% CI
≥9 vs <9	18	198	0.46	(0.23, 0.69)	0.51	(0.43, 0.58)	-0.04	(-0.29, 0.20)
≥8 vs <8	42	174	0.43	(0.28, 0.58)	0.52	(0.44, 0.60)	-0.09	(-0.26, 0.08)
≥7 vs <7	77	139	0.44	(0.33, 0.56)	0.54	(0.45, 0.63)	-0.10	(-0.24, 0.05)
≥6 vs <6	120	96	0.42	(0.33, 0.51)	0.62	(0.51, 0.73)	-0.20	(-0.34, -0.06)*
≥5 vs <5	145	71	0.44	(0.36, 0.52)	0.64	(0.52, 0.76)	-0.20	(-0.35, -0.05)*
≥4 vs <4	174	42	0.48	(0.41, 0.56)	0.61	(0.44, 0.77)	-0.13	(-0.31, 0.06)

* p<0.05
ES = effect size; CI = confidence interval; ESdiff = effect size difference

Dataset 2: EPC reports

The mean treatment effect across all studies in the EPC report dataset was 0.43 (95% CI: 0.34, 0.53). Few quality features showed differences in effect sizes according to whether these criteria were met as depicted in Table 12. A negative difference indicates that the effect size for the studies fulfilling the criterion is smaller than the effect size for the studies not meeting the criterion. The “no” and “unclear” answers were combined for all initial analyses to increase the number of analyzable studies.

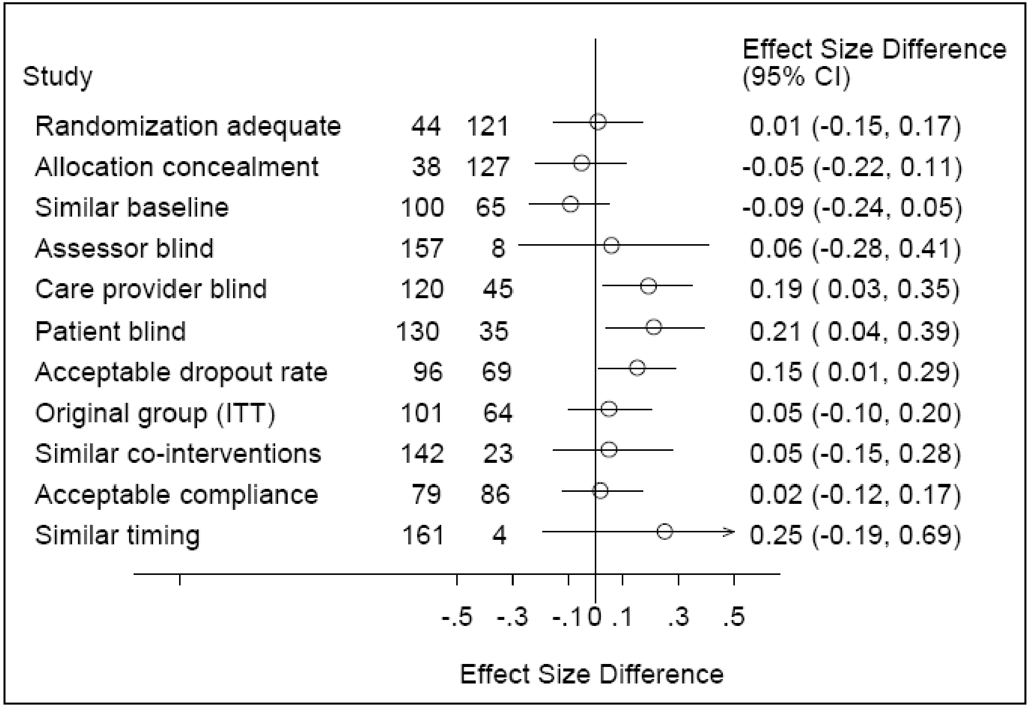
Table 12. Difference in effect sizes dataset 2 (EPC Report)

Quality Feature	Number Criterion Met	Number Criterion Not Met	Effect Size in Trials With Criterion Met		Effect Size in Trials With Criterion Not Met		Effect Size Difference	
			ES	95% CI	ES	95% CI	ESdiff	95% CI
Randomization adequate	44	121	0.44	(0.30, 0.57)	0.43	(0.34, 0.51)	0.01	(-0.15, 0.17)
Allocation concealment	38	127	0.39	(0.25, 0.53)	0.45	(0.36, 0.53)	-0.05	(-0.22, 0.11)
Similar baseline	100	65	0.40	(0.31, 0.49)	0.49	(0.37, 0.61)	-0.09	(-0.24, 0.05)
Assessor blind	157	8	0.43	(0.36, 0.51)	0.37	(0.04, 0.71)	0.06	(-0.28, 0.41)
Care provider blind	120	45	0.48	(0.40, 0.56)	0.29	(0.15, 0.43)	0.19	(0.03, 0.35)*
Patient blind	130	35	0.47	(0.40, 0.55)	0.26	(0.11, 0.42)	0.21	(0.04, 0.39)*
Acceptable dropout rate	96	69	0.50	(0.40, 0.59)	0.35	(0.24, 0.45)	0.15	(0.01, 0.29)*
Original group (ITT)	101	64	0.45	(0.36, 0.54)	0.40	(0.27, 0.52)	0.05	(-0.10, 0.20)
Similar cointerventions	142	23	0.44	(0.36, 0.52)	0.39	(0.20, 0.58)	0.05	(-0.15, 0.28)
Acceptable compliance	79	86	0.44	(0.34, 0.55)	0.42	(0.32, 0.52)	0.02	(-0.12, 0.17)
Similar timing	161	4	0.44	(0.37, 0.51)	0.19	(-0.24, 0.62)	0.25	(-0.19, 0.69)

* p<0.05
EPC = Evidence-based Practice Center; ES = effect size; CI = confidence interval; ESdiff = effect size difference;
ITT = intention to treat

High- and low-quality studies showed no difference in effect sizes for several quality features and the direction of possible bias was not consistent across dimensions. This concerned newly proposed quality features as well as established quality criteria such as blinding. For three criteria, a significant difference was found but the effect was discordant with previous studies: when care provider and patients were explicitly blinded, the average effect size in those trials was 0.48 and 0.47 compared to 0.30 and 0.27 in low or unclear quality trials. Studies that reported an acceptable dropout rate had an average effect size of 0.50; studies without description or adequate rate showed a mean effect of 0.35. Figure 10 depicts the direction of effects graphically. The difference in effect size is shown; a negative difference indicates that the high-quality studies in this dataset reported smaller effect sizes.

Figure 10. Difference in effect sizes based on quality features dataset 2 (EPC reports)



CI = confidence interval; ITT = intention to treat

Using the items constituting the Jadad scale (randomization, blinding, and dropouts), differences in effect sizes between low- and high-quality studies were 0.09 (n.s., random effects meta-regression). High-quality studies reported an effect size of 0.45 while low quality reported an effect size of 0.35 across studies. Using a fixed-effects model, the mean effect sizes were 1.07 versus 0.13 (overall effect size difference 0.94, p<0.05). The fixed-effects analysis is particularly affected by outliers. When excluding those three studies with extremely high and unmatched results, effect size differences were still 0.29. All analyses indicated that in this dataset there was no statistically significant difference in high- and low-quality studies, and often high-quality studies reported somewhat larger treatment effects, hence opposite to what we expected to find.

As shown earlier, in this dataset there was a high number of “unclear” answers. To investigate whether the combination of “no’s” and “unclear’s” may have distorted the effects of quality, we estimated the results separately, and compared the explicit negative and the unclear cases to the cases where the feature was explicitly present, that is, the criterion was clearly met, as depicted in Table 13.

Table 13. Criterion met versus not met and versus unclear (EPC reports)

Quality Feature	Met	Not	Unclear	Difference Criterion Met Versus Not Met		Difference Criterion Met Versus Unclear		Difference Criterion Met or Unclear Versus Not Met	
				ESdiff	95% CI	ESdiff	95% CI	ESdiff	95% CI
Randomization adequate	44	7	114	0.23	(-0.19, 0.65)	0.00	(-0.16, 0.16)	0.23	(-0.17, 0.64)
Allocation concealment	38	4	123	0.35	(-0.20, 0.90)	-0.07	(-0.23, 0.10)	0.40	(-0.14, 0.93)
Similar baseline	100	17	48	-0.10	(-0.35, 0.14)	-0.09	(- 0.25, 0.07)	-0.07	(-0.31, 0.16)
Assessor blind	157	2	6	-0.93	(-1.57, -0.30)*	0.40	(0.02, 0.78)*	-0.94	(-1.58, -0.30)*
Care provider blind	120	33	12	0.11	(-0.07, 0.30)	0.38	(0.11, 0.65)*	0.08	(-0.11, 0.26)
Patient blind	130	31	4	0.19	(0.01, 0.36)*	0.38	(-0.05, 0.80)	0.18	(-0.01, 0.36)
Acceptable dropout rate	96	65	4	0.14	(-0.01, 0.28)	0.33	(-0.13, 0.78)	0.13	(-0.02, 0.27)
Original group (ITT)	101	10	54	0.08	(-0.24, 0.39)	0.05	(-0.11, 0.21)	0.06	(-0.25, 0.37)
Similar co-interventions	142	5	19	-0.17	(-0.59, 0.25)	0.11	(-0.12, 0.34)	-0.18	(-0.60, 0.24)
Acceptable compliance	79	8	78	0.25	(-0.06, 0.47)	-0.01	(-0.15, 0.14)	0.26	(-0.05, 0.56)
Similar timing	161	4	0	0.25	(-0.19, 0.69)	NA	NA	0.25	(-0.19, 0.69)

* p<0.05
EPC = Evidence-based Practice Center; ESdiff = effect size difference; CI = confidence interval; ITT = intention to treat

When combining high-quality studies and studies rated as unclear, thereby giving the benefit of the doubt, the sample sizes for negative studies are very small. This stratified analysis showed that the lack of association of the feature with effect sizes cannot be generally explained by the combination of explicitly negative and unclear answers.

We also investigated in this dataset whether the conversion of dichotomous outcomes to effect sizes that was necessary in some studies may have influenced the associations between quality features and study results. Only considering original continuous outcomes, the differences between low and high quality studies ranged from -0.12 (similar baseline) to 0.33 (similar timing of outcome assessment) where a negative difference indicates that the studies with the feature showed smaller effect sizes. The 0.33 difference was based on 3 studies only where the similar outcome assessment criterion was not met or remained unclear.

Summary and Factor Scores

To explore a linear effect of quality on effect size, we again regressed the effect sizes on the total quality score values. The effect of the total quality scores weighting each item equally was negligible and not significant (0.02; p=0.233, 95% CI: -0.015, 0.062).

We also used the established factor scores that group similar items in terms of inter-item correlations as well as associations with effect sizes. Differential trends shown for the individual items as seen in the table above should become apparent using these factor empirically derived item clusters. Raw effect size (not accounting for precision with any weights) was regressed on the established factors. High-loading items from factor 1 (adequate randomization sequence, allocation concealment) were not statistically significantly associated with effect sizes (-0.172, 95% CI: -0.41, 0.06; p=0.15), nor was describing an acceptable dropout rate which mainly contributed factor 3 (0.14; 95% CI: -0.06, 0.49; p=0.16). The two blinding items and the ITT item that constitute factor 2 showed a statistically significant influence (but unexpected directionality) on effect sizes (0.24, 95% CI: 0.04, 0.44; p=0.02).

As depicted in Table 14, when comparing studies with high or low quality and applying different cutoffs, we found a marginal significant difference in effect sizes for five or more quality criteria met. However, the direction of effects was opposite to what we found in the back pain trials: high quality studies reported larger treatment effects. There was no indication that low-quality studies overestimated treatment effects; in this dataset the high-quality RCTs reported larger effects. Appendix D shows the results based on a fixed-effects model (this analysis results in smaller confidence intervals and several significant results, but the analysis is more affected by outliers).

Table 14. Comparison of different quality cutoffs using a total score (EPC reports)

Cut-off	Number Equal or Above Cutoff	Number Below Cutoff	High Quality		Low Quality		Difference	
			ES	95% CI	ES	95% CI	ESdiff	95% CI
≥9 vs <9	42	123	0.54	(0.40, 0.67)	0.39	(0.31, 0.47)	0.15	(-0.01, 0.30)
≥8 vs <8	65	100	0.45	(0.34, 0.56)	0.42	(0.32, 0.51)	0.03	(-0.11, 0.18)
≥7 vs <7	103	62	0.43	(0.35, 0.52)	0.43	(0.31, 0.55)	0.01	(-0.14, 0.16)
≥6 vs <6	135	30	0.46	(0.38, 0.53)	0.30	(0.12, 0.47)	0.16	(-0.03, 0.35)
≥5 vs <5	149	16	0.45	(0.38, 0.53)	0.18	(-0.06, 0.42)	0.27	(0.02, 0.52)*
≥4 vs <4	160	5	0.44	(0.37, 0.51)	0.04	(-0.39, 0.46)	0.41	(-0.02, 0.84)

* p<0.05
EPC = Evidence-based Practice Center; ES = effect size; CI = confidence interval; ESdiff = effect size difference

The above table also clearly demonstrates the imbalance of this dataset. There were very few low-quality studies in this dataset. When comparing studies that only reached 4 out of 11 possible quality scores, only 5 studies could be included in the analysis.

Dataset 3: Published “Pro-bias” Dataset

Given the discrepant results between the analyses of the back pain dataset and the EPC reports dataset, we decided to add a third dataset. We were struck, in particular, by the observation that in the EPC dataset for the majority of quality features we actually found results in the opposite direction as expected from prior research (high-quality studies reported larger effect sizes). As outlined in the method section for our third dataset, we therefore determined that we should use one where established criteria such as the Jadad and Schulz items had known values in the expected direction. For that reason, we decided to use a replication of the dataset used by Moher and colleagues in their original validation of these quality features.

As opposed to the prior two datasets, all the outcomes in this dataset used dichotomous outcomes. Therefore, instead of a difference in effect sizes we use, as Moher and colleagues did in their original analysis, the odds ratio as the measure of effect and the ratio of odds as assessment of the effect of a quality criterion across studies. The overall odds ratio across studies was 0.47 (95% CI: 0.42, 0.52). An odds ratio below 1 indicates the treatment group is doing better than the control.

In this dataset we also applied the full Jadad scale and the criteria proposed by Schulz (1995), which included concealment of allocation using the original scoring instructions. We compared adequate randomization (score=2 versus <2), blinding (score=2 versus <2) and withdrawals (score 1 versus 0) and a total score of 3 or more (out of 5) compared to less than 3 for the total Jadad score. For the Schulz scores, we compared criterion met versus not met or unclear. Table 15 shows the results for these established quality criteria.

Table 15. Difference in odds ratios for Jadad and Schulz criteria

Quality Feature	Number Criterion Met	Number Criterion Not Met	Effect Size in Trials With Criterion Met		Effect Size in Trials With Criterion Not Met		Effect Size Difference	
			OR	95% CI	OR	95% CI	ROR	95% CI
Random Effects Meta-regression								
Jadad: randomization=2	33	67	0.48	(0.36, 0.57)	0.45	(0.35, 0.57)	0.95	(0.62, 1.44)
Jadad: blinding=2	36	64	0.43	(0.31, 0.61)	0.47	(0.37, 0.60)	1.08	(0.72, 1.64)
Jadad: withdrawals=1	74	26	0.48	(0.38, 0.60)	0.41	(0.29, 0.60)	0.87	(0.56, 1.34)
Jadad: total≥3	62	38	0.46	(0.36, 0.60)	0.45	(0.33, 0.61)	0.97	(0.65, 1.45)
Schulz: concealment	26	74	0.58	(0.40, 0.86)	0.42	(0.34, 0.53)	0.72	(0.46, 1.13)
Schulz: sequence	30	70	0.45	(0.31, 0.65)	0.46	(0.36, 0.58)	1.01	(0.66, 1.56)
Schulz: analysis	41	59	0.50	(0.37, 0.68)	0.43	(0.33, 0.55)	0.85	(0.57, 1.26)
Schulz: blinding	66	34	0.44	(0.34, 0.56)	0.50	(0.35, 0.69)	1.13	(0.74, 1.71)
Fixed Effects Model								
Jadad: randomization=2	33	67	0.51	(0.43, 0.61)	0.45	(0.40, 0.51)	0.88	(0.70, 1.09)
Jadad: blinding=2	36	64	0.45	(0.37, 0.54)	0.47	(0.42, 0.53)	1.05	(0.85, 1.31)
Jadad: withdrawals=1	74	26	0.52	(0.46, 0.59)	0.39	(0.33, 0.46)	0.74	(0.60, 0.92)*
Jadad: total≥3	62	38	0.51	(0.44, 0.58)	0.43	(0.37, 0.49)	0.85	(0.69, 1.03)
Schulz: concealment	26	74	0.57	(0.46, 0.69)	0.44	(0.39, 0.49)	0.77	(0.61, 0.97)*
Schulz: sequence	30	70	0.50	(0.41, 0.60)	0.46	(0.41, 0.51)	0.92	(0.74, 1.15)
Schulz: analysis	41	59	0.51	(0.44, 0.59)	0.43	(0.38, 0.50)	0.85	(0.70, 1.04)
Schulz: blinding	66	34	0.48	(0.42, 0.55)	0.44	(0.38, 0.52)	0.92	(0.75, 1.12)

* p<0.05
OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios

As the original paper by Moher and colleagues does not specify whether they used a fixed-effects or a random-effects model, we present the results in Table 15 using both methods. Our fixed-effect results come closest to the original results presented by Moher et al. (1998). The dimensions overall showed consistent results, with the ratio of odds ratios (ROR) for low-quality studies compared to high-quality studies being less than 1, meaning high-quality studies reported smaller treatment effects (i.e., larger ORs) than did low-quality studies. Using a fixed-effects model, the concealment criterion item of Schulz is statistically significantly associated with smaller treatment effects, and the Jadad scale is nearly so (ROR = 0.85, 95% CI: 0.69, 1.03). Using the random effects model, both point estimates go in the expected direction, but neither result is statistically significantly different from an ROR of 1.

Table 16 shows the odds ratios of studies where the criterion is met and odds ratios for studies where the criterion is not met, either because of poor reporting or due to the design, conduct, or analysis of the individual study, using a fixed-effects model. The corresponding results using a meta-regression model assuming random effects are shown in Appendix E. To assess the difference between these two study types, we estimated an ROR. Again, an ROR less than 1 indicated that the studies that did not meet the quality criteria had a better treatment effect than those studies that met the quality criteria.

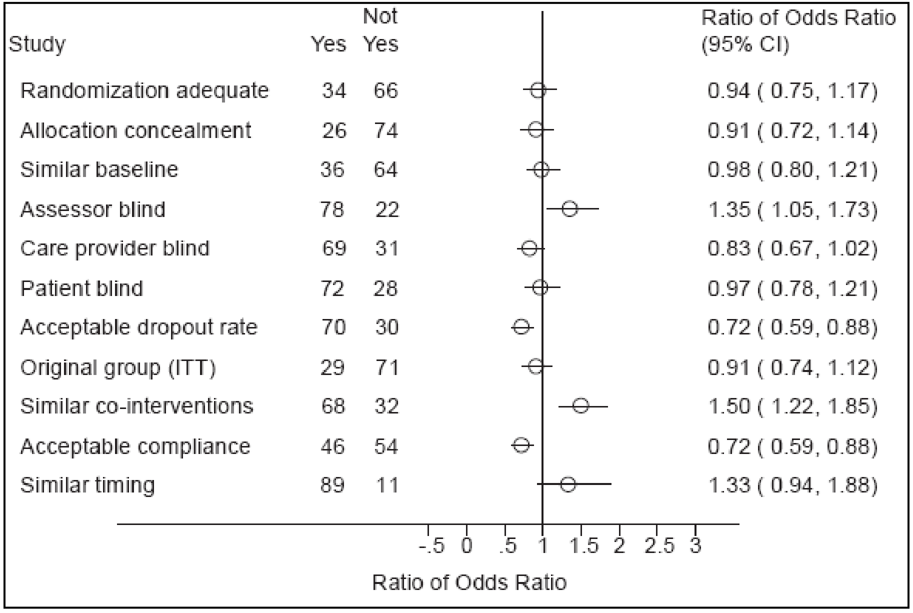
Table 16. Difference in odds ratios for proposed quality criteria dataset 3 (“pro-bias”)

Quality Feature	Number Criterion Met	Number Criterion Not Met	Effect Size in Trials With Criterion Met		Effect Size in Trials With Criterion Not Met		Effect Size Difference	
			OR	95% CI	OR	95% CI	ROR	95% CI
Randomization adequate	34	66	0.49	(0.41, 0.59)	0.46	(0.41, 0.52)	0.94	(0.75, 1.17)
Allocation concealment	26	74	0.50	(0.41, 0.62)	0.46	(0.41, 0.51)	0.91	(0.72, 1.14)
Similar baseline	36	64	0.47	(0.40, 0.56)	0.46	(0.41, 0.52)	0.98	(0.80, 1.21)
Assessor blind	78	22	0.44	(0.39, 0.49)	0.59	(0.47, 0.74)	1.35	(1.05, 1.73)
Care provider blind	69	31	0.50	(0.41, 0.57)	0.41	(0.35, 0.49)	0.83	(0.67, 1.02)
Patient blind	72	28	0.47	(0.42, 0.53)	0.46	(0.38, 0.55)	0.97	(0.78, 1.21)
Acceptable dropout rate	62	38	0.54	(0.47, 0.62)	0.39	(0.34, 0.46)	0.72	(0.59, 0.88)
Original group (ITT)	29	71	0.49	(0.42, 0.58)	0.45	(0.40, 0.51)	0.91	(0.74, 1.12)
Similar cointerventions	68	32	0.40	(0.35, 0.46)	0.60	(0.51, 0.71)	1.50	(1.22, 1.85)
Acceptable compliance	46	54	0.56	(0.48, 0.66)	0.41	(0.36, 0.46)	0.72	(0.59, 0.88)
Similar timing	89	11	0.45	(0.41, 0.50)	0.60	(0.43, 0.84)	1.33	(0.94, 1.88)

OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios; ITT = intention to treat

The direction of effects for the 11 quality features was not uniform but for the majority of quality domains the low quality studies reported smaller odds ratios thereby overestimating the treatment effect. Differences between low- and high-quality studies were statistically significant for the quality items assessor blinding, acceptable dropout rate, similar cointerventions, and acceptable compliance. Figure 11 below displays the effects graphically.

Figure 11. Ratio of odds ratio based on quality features dataset 3 (“pro-bias”), FE



FE = based on fixed-effects model; CI = confidence interval; ITT = intention to treat

Summary and Factor Scores

We also computed total quality scores for each study based on the 11 assessed quality features. Using this summary score and regressing effect size on quality, thereby assuming a linear relationship between the two variables, we find no statistically significant effect (estimate = 0.033, p=0.486, 95% CI: -0.061, 0.127) for total quality, regardless of the employed meta-regression model.

Using the factor scores that group related items, in terms of intercorrelations as well as in their strength of association with study results (the MIMIC model) we can show differences in groups of quality measures. The regression effect of log odds ratios on the randomization factor (high loadings in this dataset: adequate randomization sequence generation, concealed treatment allocation; Factor 1) was 0.05, and was not significant (p=0.792). The effect of the blinding factor (high loadings in this dataset: provider blind, patient blind, assessor blind) was also not statistically significant (-0.06, p = 0.71), again regardless of the employed method. Factor 3 (high loadings in this dataset: acceptable dropout rate, similar baseline, original group (ITT), similar cointerventions, and similar timing) had a marginally nonsignificant, and negative effect on the log odds ratios (-0.22, p=0.07).

When comparing studies with high or low quality, and applying different cutoffs, and applying the proposed quality criteria of the Cochrane back review group to compute a total score, we find that based on a fixed-effects model, a cutoff of 5 or 6 differentiates the studies statistically significant as shown in Table 17.

Table 17. Comparison of different quality cutoffs using a total score dataset 3 (“pro-bias”)

Cut-off	Number Equal or Above Cutoff	Number Below Cutoff	High Quality		Low Quality		Difference	
			OR	95% CI	OR	95% CI	ROR	95% CI
≥9 vs <9	14	86	0.43	(0.32, 0.57)	0.47	(0.42, 0.52)	1.09	(0.81, 1.46)
≥8 vs <8	26	74	0.53	(0.43, 0.64)	0.45	(0.40, 0.50)	0.85	(0.68, 1.07)
≥7 vs <7	44	56	0.45	(0.39, 0.53)	0.48	(0.42, 0.54)	1.05	(0.86, 1.29)
≥6 vs <6	62	38	0.52	(0.45, 0.59)	0.40	(0.34, 0.47)	0.77	(0.63, 0.95)*
≥5 vs <5	76	24	0.50	(0.44, 0.56)	0.39	(0.32, 0.48)	0.79	(0.63, 0.99)*
≥4 vs <4	86	14	0.48	(0.43, 0.54)	0.40	(0.32, 0.50)	0.83	(0.65, 1.07)

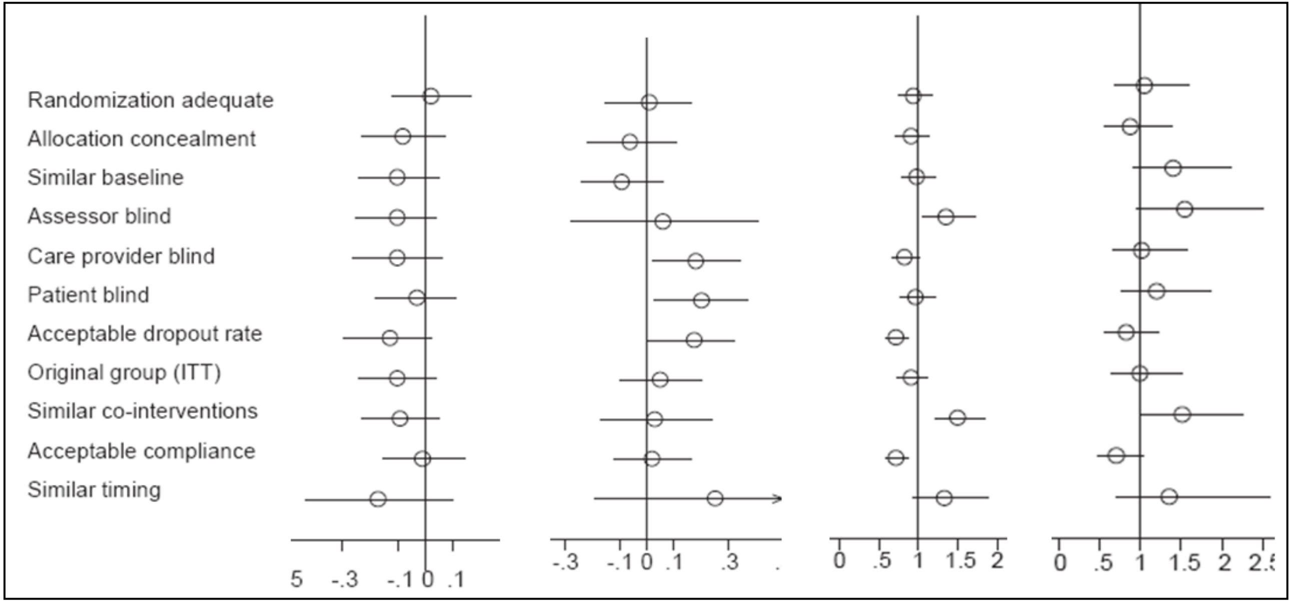
* p<0.05
OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios

High-quality studies show less pronounced treatment effects compared to low-quality studies. The equivalent analysis using a random effects meta-regression is reported in Appendix E. Using this model, a cutoff of 7 differentiates high- and low-quality studies best, but this result does not reach statistical significance.

Comparison Across Datasets

Figure 12 allows a comparison of observed indicators of bias for each individual quality feature across datasets.

Figure 12. Differences in effect sizes across datasets



Left to right: back pain data, EPC reports, “pro-bias” (fixed-effects model), “pro-bias” (random-effects model)
ITT = intention to treat

While the back pain dataset shows small but consistent results across quality criteria indicating that studies fulfilling the quality criterion report smaller effect sizes, the EPC dataset indicate for the majority of quality dimensions that high-quality studies reported larger treatment effects. The third dataset shows the most variation across quality criteria. In the fixed-effects analysis, the differences across high- and low-quality studies reach statistical significance.

Only allocation concealment showed consistent results across datasets. Unconcealed trials reported larger effect sizes in the back pain dataset (effect size difference in random effects meta-regression -0.08 (95% CI: -0.23, 0.07) and the EPC report dataset (effect size difference in random effects meta-regression -0.06 (95% CI: -0.22, 0.11), and the ratio of odds ratios in the third dataset also showed that unconcealed trials reported larger treatment effects (ROR=0.91, 95% CI: 0.72, 1.14).This analysis indicates that unconcealed trials tend to overestimate the treatment effect.

Across all three datasets, there was no statistically significant linear effect of quality on effect sizes. Regression models could not show that the effect size decreased linearly with increasing total quality scores.

The factors derived through factor analysis showed no statistically significant association between quality and effect sizes. One exception was the blinding/ITT factor (factor 2 in the EPC report data), but here with unexpected directionality (larger effects observed in high-quality studies).

Comparing different cutoffs shows that five or more fulfilled criteria differentiate high- and low-quality studies best across datasets. In dataset 1 (back pain), the difference between effect sizes was -0.20 for both, five criteria fulfilled or more (95% CI: -0.34, -0.06), and six criteria fulfilled or more (95% CI: -0.35, -0.05). In the third dataset, the ratio of odds ratios was 0.79 (95% CI: 0.63, 0.95) and 0.77 (95% CI: 0.63, 0.99) respectively for five and six criteria met (based on a fixed-effects model). In both cases, low-quality studies overestimated treatment effects. However, in the EPC report dataset, a cutoff of five or more quality criteria met also resulted in a statistically significant result (effect size difference 0.27, 95% CI: 0.02, 0.52), but the direction was opposite to our expectations. Low-quality studies did not overestimate treatment effects but reported smaller effect sizes than high-quality trials in this dataset.

Why the Association Between Quality Features and Effect Sizes Might Vary Across Datasets: Moderators and Confounders

As seen above, the association between quality features and effect sizes varies across the three employed datasets. In two of the datasets, the Jadad and Schulz criteria show associations that are consistent with that found by others, namely that higher quality studies have smaller estimates of effect. In these two datasets, the CBRG internal validity items also, in general, show the predicted relationships between quality and effect size, and in each a summary score of the 11 items is useful for distinguishing high- and low-quality trials (cutoff 5 to 7 quality items met). However, in the EPC dataset, the majority of quality features show either no relationship or a paradoxical relationship with effect size. To try and understand why these differences might exist, we undertook an analysis looking at moderators and confounders based on our conceptual models outlined in the method section.

We wanted to investigate the effect of the size of the treatment effect, the effect of the condition being treated, the effect of the type of outcome, and the effect of the observed variance in quality features.

Variable 1: Size of Treatment Effect

One variable we pursued was the size of the treatment effect reported in each individual meta-analysis. Our hypothesis was that if a treatment is very effective, this may minimize any associations between quality and outcomes. Depending on the type of intervention, the achieved effect can vary systematically across studies, thereby possibly confounding an effect of the association of quality and study results across studies. We added the treatment effect observed in each individual meta-analysis to the regression model.

Table 18 shows the differences between low- and high-quality studies for the EPC report dataset when controlling for the mean treatment effect in each meta-analysis.

Table 18. Controlling for size of treatment effect dataset 2 (EPC reports)

Quality Feature	Number Criterion Met	Number Criterion Not Met	Effect Size in Trials With Criterion Met		Effect Size in Trials With Criterion Not Met		Effect Size Difference	
			ES	95% CI	ES	95% CI	ESdiff	95% CI
Randomization adequate	44	121	0.41	(0.31, 0.51)	0.41	(0.35, 0.48)	0.00	(-0.12, 0.11)
Allocation concealment	38	127	0.40	(0.29, 0.50)	0.42	(0.36, 0.48)	-0.02	(-0.14, 0.10)
Similar baseline	100	65	0.40	(0.31, 0.46)	0.44	(0.35, 0.53)	-0.04	(-0.16, 0.07)
Assessor blind	157	8	0.41	(0.35, 0.46)	0.53	(0.28, 0.79)	-0.13	(-0.39, 0.13)
Care provider blind	120	45	0.40	(0.34, 0.47)	0.44	(0.33, 0.55)	-0.03	(-0.17, 0.10)
Patient blind	130	35	0.41	(0.35, 0.47)	0.42	(0.30, 0.55)	-0.01	(-0.16, 0.13)
Acceptable dropout rate	96	69	0.41	(0.34, 0.48)	0.42	(0.33, 0.50)	0.00	(-0.12, 0.11)
Original group (ITT)	101	64	0.41	(0.34, 0.47)	0.43	(0.33, 0.52)	-0.02	(-0.14, 0.10)
Similar cointerventions	142	23	0.41	(0.35, 0.47)	0.43	(0.28, 0.57)	-0.02	(-0.18, 0.14)
Acceptable compliance	79	86	0.40	(0.33, 0.48)	0.42	(0.35, 0.50)	-0.02	(-0.13, 0.09)
Similar timing	161	4	0.42	(0.36, 0.47)	0.30	(-0.02, 0.62)	0.12	(-0.21, 0.44)

ES = effect size; CI = confidence interval; ESdiff = effect size difference; ITT = intention to treat

There is no indication that controlling for treatment effect size reveals associations of quality and effect sizes. In fact, controlling for this variable eliminates differences between high- and low-quality studies, effect size differences range around zero. The differential effect of possible bias (sometimes indicating that low-quality studies show larger effect sizes than high-quality studies, sometimes indicating that high-quality studies show larger effect sizes) that characterizes the EPC report dataset appears to be primarily based on this treatment effect variable.

A similar result was observed in the third dataset, which showed similar results overall to the original back pain results. The differences in effect sizes comparing high- and low-quality studies are negligible for several quality domains with the exception of reported provider blinding, acceptable dropout rate and the ITT item (original group) when comparing for size of treatment effect, as shown in Table 19.

Table 19. Controlling for size of treatment effect dataset 3 (“pro-bias”)

Quality Feature	Number Criterion Met	Number Criterion Not Met	OR in Trials With Criterion Met		OR in Trials With Criterion Not Met		OR Difference	
			OR	95% CI	OR	95% CI	ROR	95% CI
Randomization adequate	34	66	0.44	(0.31, 0.63)	0.46	(0.37, 0.59)	1.02	(0.72, 1.43)
Allocation concealment	26	74	0.42	(0.31, .0.58)	0.44	(0.37, 0.52)	1.04	(0.72, 1.49)
Similar baseline	36	64	0.44	(0.34, 0.57)	0.44	(0.36, 0.53)	0.99	(0.72, 1.35)
Assessor blind	78	22	0.44	(0.37, 0.52)	0.44	(0.31, 0.62)	0.99	(0.67, 1.47)
Care provider blind	69	31	0.48	(0.40, 0.57)	0.36	(0.28, 0.47)	0.75	(0.55, 1.02)
Patient blind	72	28	0.44	(0.37, 0.520	0.43	(0.33, 0.58)	0.99	(0.70, 1.39)
Acceptable dropout rate	62	38	0.50	(0.41, 0.61)	0.38	(0.30, 0.47)	0.75	(0.56, 1.00)
Original group (ITT)	29	71	0.53	(0.41, 0.67)	0.39	(0.33, 0.47)	0.74	(0.55, 1.01)
Similar cointerventions	68	32	0.42	(0.35, 0.51)	0.47	(0.36, 0.61)	1.12	(0.81, 1.55)
Acceptable compliance	46	54	0.46	(0.37, 0.58)	0.42	(0.34, 0.51)	0.91	(0.67, 1.24)
Similar timing	89	11	0.44	(0.37, 0.52)	0.42	(0.25, 0.70)	0.96	(0.56, 1.64)

OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios; ITT = intention to treat

Using summary quality scores and regressing study results on quality, we find in both new datasets no significant results that indicate a significant linear relationship between these two variables. When controlling for the mean treatment effect of each of the 12 meta-analyses in the EPC reports dataset, the regression results are still -0.02 (p=0.38; 95% CI: -0.06, 0.02) as opposed to 0.02. When controlling for the mean effect size of each of the 11 meta-analyses in the third dataset, results are also unchanged: 0.04 (p=0.376; 95% CI: -0.05, 0.13), previously 0.03.

Variable 2: Condition Being Treated

In the EPC report dataset, we found no clear associations between quality and effect size across all studies. Hence, we wanted to investigate whether pooling across meta-analyses masks associations between the quality features and effect sizes. In order to see whether the associations between quality and effect sizes are consistent or notably different across clinical fields, we stratified the studies by the condition being treated or the clinical field. Only this dataset was considered suitable for this analysis (see method section).

Table 20 shows the effect size difference for high- (criterion fulfilled) and low- (criterion not fulfilled) quality studies for each meta-analysis individually for the EPC report dataset studies. Each cell had to have at least three trials with the feature present versus absent or unclear to be estimated.

Table 20. Effect size differences studies fulfilling criterion versus not by clinical field (EPC reports)

Quality Feature	Alzheimer's	Arthritis	CDSM	Chromium	Epilepsy	Glucosamine	OCD	Omega 3	Orlistat	SAMe	SMBG	Vitamin E
Randomization adequate	.04	n/a	-0.08	n/a	.00	-.10	-.12	-.01	.32	-.03	n/a	-.17
Allocation concealment	.07	n/a	-0.26	n/a	.00	-.19	-.53	.01	.32	n/a	n/a	.3
Similar baseline	.08	n/a	-0.04	-.47	-.11	-.20	.28	-.10	n/a	-.53	.18	-.41
Assessor blind	n/a	n/a	n/a	n/a	n/a	n/a	n/a	-.02	n/a	n/a	n/a	.36
Care provider blind	n/a	n/a	n/a	n/a	n/a	.14	n/a	.00	n/a	n/a	n/a	.14
Patient blind	n/a	n/a	n/a	n/a	n/a	n/a	n/a	-.06	n/a	n/a	n/a	n/a
Acceptable dropout rate	-.12	n/a	-0.03	.36	-.05	.09	n/a	.10	.01	-1.39	-.01	.38
Original group (ITT)	n/a	n/a	0.31	-.45	-.04	-.12	.12	-.02	-.04	-.85	-.05	.25
Similar cointerventions	n/a	n/a	0.04	n/a	.17	-1.02	.79	-.23	n/a	n/a	.13	n/a
Acceptable compliance	n/a	n/a	0.01	.45	-.03	-.37	.18	-.15	-.02	.17	.1	-.02
Similar timing	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	.22	n/a

EPC = Evidence-based Practice Center; n/a = not available (fewer than three trials in each group); ITT = intention to treat; CDSM: chronic disease self-management; OCD: obsessive-compulsive-disorder; SAMe: S-adenosylmethionine; SMBG: self-monitoring of blood glucose

The table shows that in each meta-analysis, many comparisons could not be computed due to lack of variance across studies in these smaller units. Very often there was not enough information to judge the criterion, meaning fewer than three studies in each meta-analysis scored definitely positive. For some quality features (e.g., was the timing of outcome assessment similar in both groups) there were no studies not meeting this criterion, so again the difference in effect sizes could not be computed.

There was no clear support for the hypothesis of the condition being treated masking the associations of quality through pooling. The quality effects were not confounded as outlined in the method section; the pooling appeared to cancel out conflicting effects across fields resulting in the negligible pooled effects seen. The effect of the quality features varied across clinical fields in that a quality criterion that was met sometimes indicated smaller effect sizes and sometimes larger effect sizes. For example, in most fields, the adequacy of randomization sequence showed a small difference between high- and low-quality studies in the direction that the high-quality studies reported smaller effect sizes, but the direction of effect was reversed for the orlistat trials. Most consistency in the expected effect (smaller effect sizes when quality criterion met) was found for randomization sequence adequate, similar baseline, ITT analysis, and acceptable compliance.

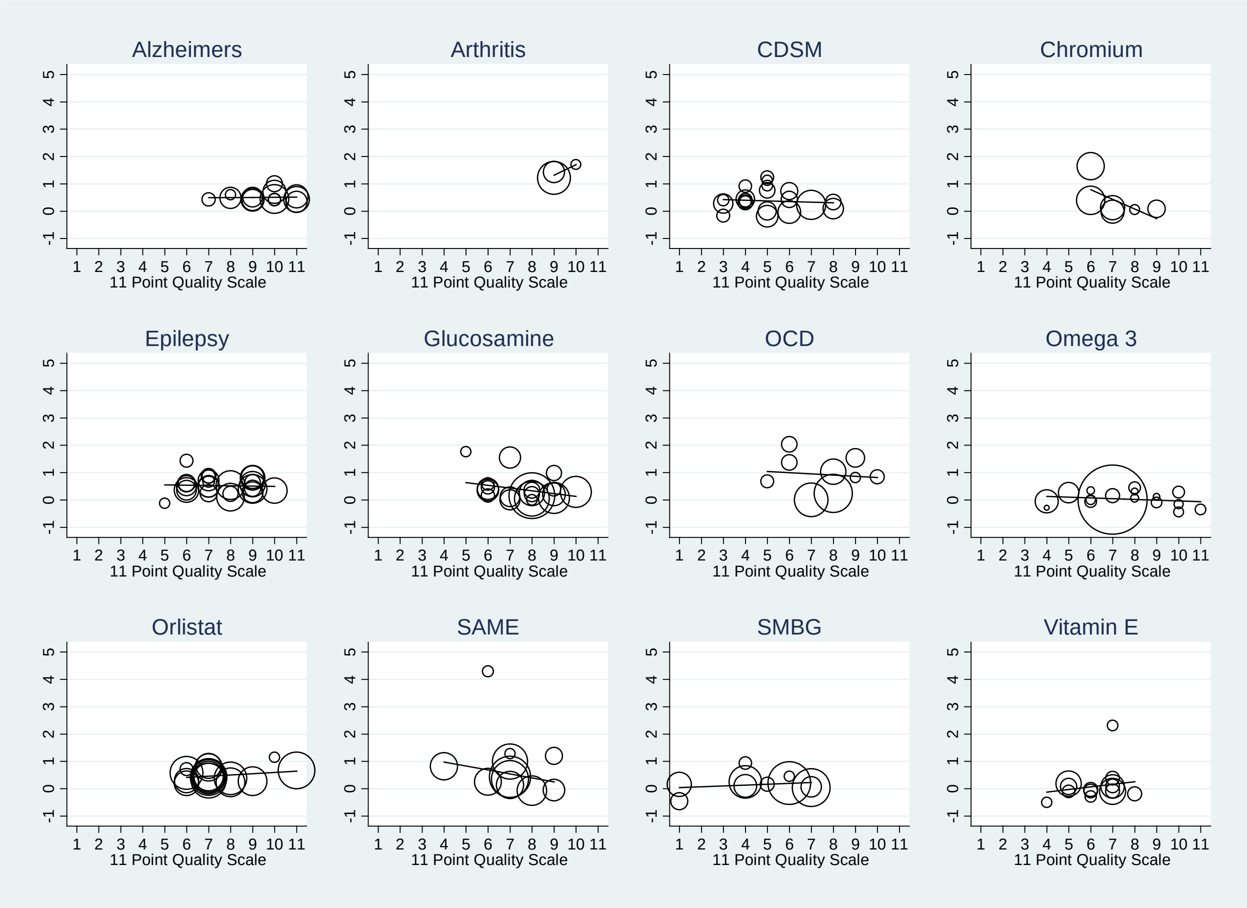
In addition, we repeated this analysis using a total quality score that considered all assessed quality features. Again, it is possible that the type of condition acts as a moderator or confounder. In some studies quality may have an effect on study results, but these effects are masked by other studies where there is no association between quality and study results. The effect of the total quality sum score on study results was calculated for each condition. The meta-regression slopes for each condition are shown in Table 21 and Figure 13.

Table 21. Total quality regressed on effect size (EPC reports)

Condition	Estimate	SE
1. Alzheimer's	0.01	0.08
2. Arthritis	0.38	0.40
3. CDSM	-0.02	0.06
4. Chromium	-0.35	0.18
5. Epilepsy	-0.01	0.06
6. Glucosamine	-0.10	0.07
7. OCD	-0.04	0.07
8. Omega 3	-0.03	0.04
9. Orlistat	0.05	0.06
10. SAMe	-0.14	0.12
11. SMBG	0.03	0.06
12. Vitamin E	0.09	0.12

EPC = Evidence-based Practice Center; SE = standard error; CDSM: chronic disease self-management; OCD: obsessive-compulsive-disorder; SAMe: S-adenosylmethionine; SMBG: self-monitoring of blood glucose

Figure 13. Meta-regression slopes showing relationship between total quality and effect size in each type of study



Circle is proportional to the size of the study; the line represents the effect on effect size.

The interaction effect of condition by total quality was not statistically significant (p=0.574) , meaning that slopes are not significantly different from each other indicating that condition is not a moderator (confounder) of the association between quality and effect sizes.

Variable 3: Type of Outcome

Table 22 shows differences of studies fulfilling a quality criterion compared to studies not meeting the criterion when controlling for the type of outcome, that is, objective or less prone to measurement error, versus other outcomes. In dataset 2 (EPC reports), 47 studies were classified as having an objective outcome as opposed to other endpoints; in the third dataset there were 35 studies (the back pain dataset did not include objective outcomes).

Table 22. Difference in effect sizes between high- and low-quality studies, controlled for type of outcome

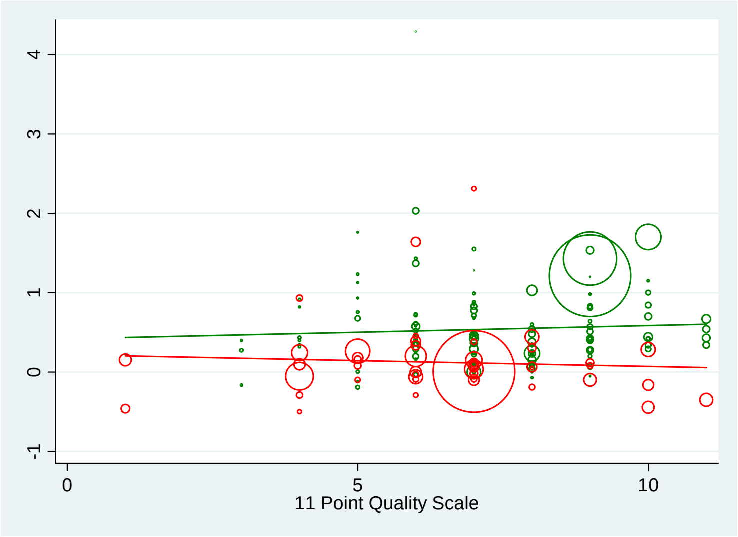
Quality feature	EPC reports		Published dataset	
	ESdiff	95% CI	ROR	95% CI
Randomization adequate	0.01	(-0.14, 0.15)	1.12	(0.74, 1.70)
Allocation concealment	-0.05	(-0.20, 0.10)	1.02	(0.65, 1.60)
Similar baseline	-0.01	(-0.15, 0.12)	1.03	(0.67, 1.59)
Assessor blind	-0.06	(-0.38, 0.25)	1.27	(0.78, 2.06)
Care provider blind	-0.14	(-0.32, 0.05)	0.71	(0.46, 1.10)
Patient blind	-0.16	(-0.37, 0.05)	0.66	(0.40, 1.08)
Acceptable dropout rate	0.08	(-0.05, 0.21)	0.83	(0.56, 1.22)
Original group (ITT)	-0.10	(-0.24, 0.05)	0.91	(0.60, 1.37)
Similar cointerventions	-0.05	(-0.24, 0.13)	1.24	(0.82, 1.89)
Acceptable compliance	-0.02	(-0.15, 0.11)	0.76	(0.52, 1.12)
Similar timing	0.09	(-0.32, 0.49)	0.74	(0.36, 1.51)

EPC = Evidence-based Practice Center; ESdiff = effect size difference; CI = confidence interval; ROR = ratio of odds ratios; ITT = intention to treat

For most individual quality domains differences between high- and low-quality studies were not more pronounced when controlling for the type of outcome. Controlling for the type of outcome, the association between a total quality score and reported effect size is 0.01 (SE 0.02) in the EPC reports and 0.02 (SE 0.05) in the “pro-bias” dataset.

In a moderator analysis, an interaction effect between total quality and the type of outcome measure was investigated. The slope of total quality for studies with nonobjective outcomes in the EPC report sample was 0.02 (95% CIs -0.03, 0.06; p=0.446). For studies with objective outcomes, the slope was -0.01 (95% CIs: -0.08, 0.05; p=0.639). The difference in slopes was not statistically significant (0.03; 95% CIs: -0.04, 0.10; p=0.411). Figure 14 shows the difference between these effects.

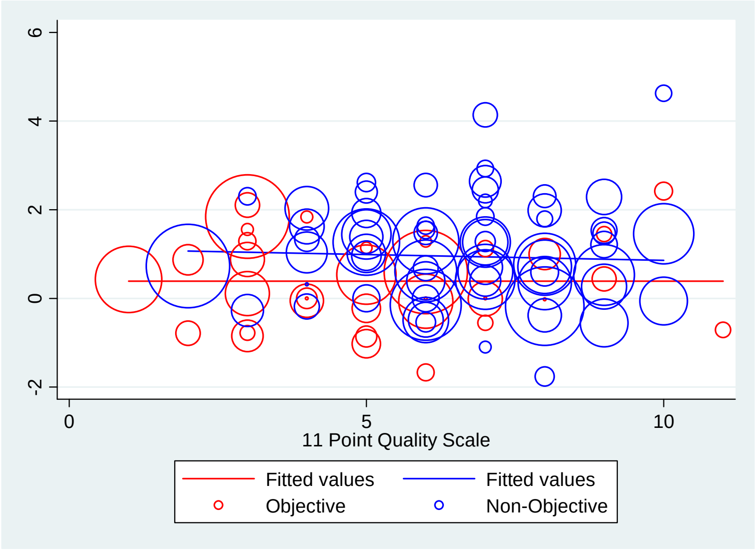
Figure 14. Association quality – effect size, moderator type of outcome (EPC reports)



Green circles and line represent objective outcomes, red represent nonobjective outcomes.

In the third dataset, the slope for nonobjective studies was -0.026 (SE 0.06; p=0.686); the slope for objective studies was 0.00 (SE 0.08; p=0.997). The interaction effect, the difference between these slopes was not statistically significant (0.026; SE 0.10; p=0.80). Figure 15 graphically depicts the effects of the type of outcome as a moderator, indicating a slightly different, that is, not parallel slope for the two types of studies.

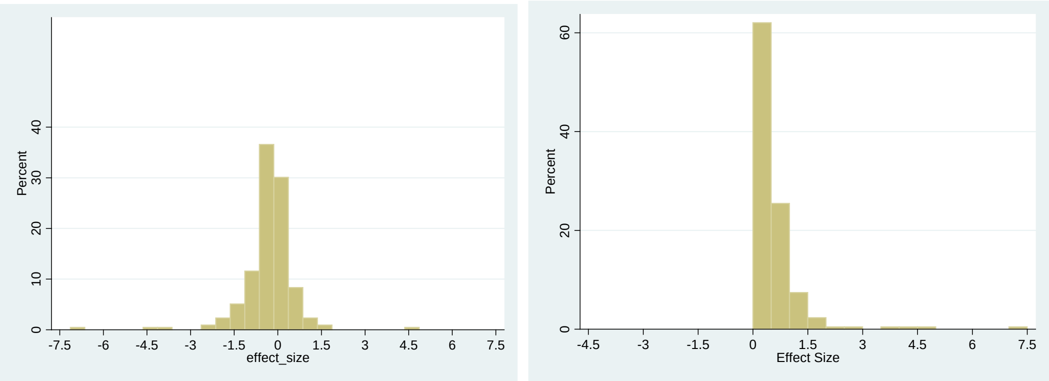
Figure 15. Association quality – effect size, moderator type of outcome ("pro-bias")



Variable 4: Variance in Effect Sizes

In order to test the hypothesis that the variance in effect sizes explains differences between datasets, we compared the effect size distribution across datasets. Figures 16 and 17 show the distribution of the effect sizes reported for each included study in the three datasets.

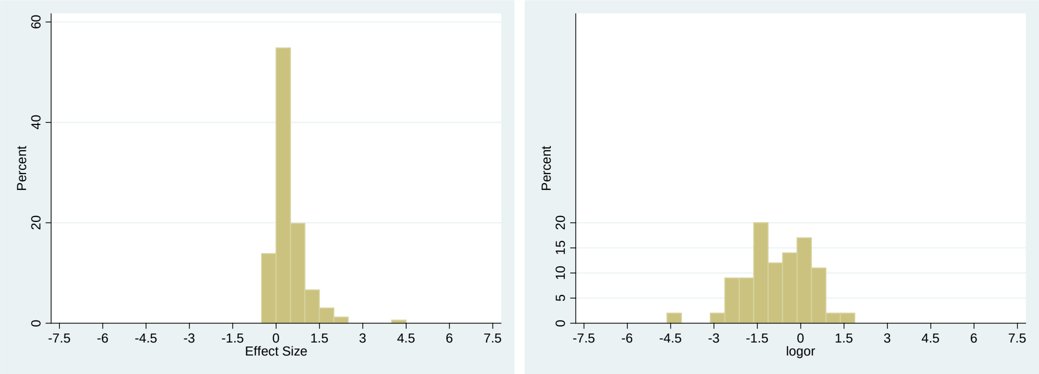
Figure 16. Effect size distribution dataset 1



Back pain data, absolute values on the right

The distribution of effect sizes in the original back pain articles was approaching a normal distribution. The analyses presented in this report are based on absolute values but the quality effect size association results are very similar when using the original reported effect sizes. The effect size distribution of the other publications is shown in Figure 17.

Figure 17. Effect size distribution dataset 2 (EPC reports) and dataset 3 ("pro-bias")



The three datasets show different distributions of effect sizes. The back pain and the published "pro-bias" data stem from datasets with substantial variation in reported effect sizes and approaching a symmetric distribution. The effect size distribution in the EPC report dataset was restricted; most reported results were small and very few were negative.

Discussion

This report quantifies the risk of bias associated with selected quality criteria across three different datasets. Our analyses show that the association between quality features and effect sizes is complex and may vary according to factors yet to be explored.

Quality of the Reporting

We found in all datasets that the quality of the reporting was lacking. Many studies in our datasets reported insufficient information to know whether a quality feature was met or not. Poor reporting in publications does not necessarily mean poor study quality. The majority of our analyses compared studies reporting positive features with studies where the quality feature was not reported, such as concealed treatment allocation. Hence, we were primarily concerned with demonstrating the effects of high-quality studies that reported a feature, and the expression of the feature was an indicator of high quality. Here, quality included the reporting, as well as quality in the design, conduct, and analysis of the study. Similarly, a recent study by Hartling, Ospina, Liang, et al. (2009) applying the Cochrane risk of bias (2008 version) compared studies with a high risk of bias with studies of low quality or unclear quality. We could also show in the analyzed datasets that the reporting has improved since publication of the Consort statement, in accordance with the observation of other reviewers (Kane, Wang, and Garrard, 2007).

However, it has to be considered that all included randomized controlled trials within datasets were identified through meta-analyses. The trials were all considered adequate for inclusion in a published meta-analysis. Moja et al. (2005) showed that 12 percent of Cochrane and the Database of Abstracts of Reviews of Effects (DARE) selected reviews used quality as an inclusion criterion. For Evidence-based Practice Center (EPC) reports in particular, this approach is also not uncommon, especially when sufficient high-quality studies are available. The quality of randomly selected trials may be lower still than encountered in our selected datasets.

Psychometric Analysis

First, our psychometric analysis indicates that the quality criterion “similar timing of outcome assessment” should be reassessed for inclusion in the 11-item list. This criterion is usually met; only 2 to 11 percent of studies across datasets indicate the possibility or evidence of differential outcome assessment.

Furthermore, we explored the interrelationship between quality features psychometrically and through the use of a multiple indicator multiple cause (MIMIC) model factor analysis. We were able to show that across all three datasets, several individual quality features showed substantial intercorrelations but the complete set of 11 items did not show marked internal consistency.

Although conceptually presumably independent, in practice we find that studies that observe good practice for selected quality features often also do so for other features. Studies that reported an adequate method of randomization sequence generation tended to also report the use of adequate treatment allocation concealment (intercorrelations ranging from 0.49 to 0.74 across datasets). Furthermore, the Cochrane Back Review Group (CBRG) criteria list differentiates patient, provider, and assessor blinding, but our analyses did not provide support for this differentiation. In our empirical data samples these features are substantially intercorrelated, and in particular the provider blinding does not appear to contribute unique information. In two of the datasets, one of the three blinding items had to be excluded from the factor analyses due to collinearity. Therefore, while appealing conceptually, the distinction between blinding the patient, provider, or outcome assessor may not all independently assess study quality.

However, when treating the items as the indicator of the same underlying construct, we found insufficient internal consistency to indicate a homogenous quality construct. The Cronbach's alpha values in all three datasets were below values expected for a psychometric scale (alphas ranging from 0.55 to 0.61).

A factor analysis taking into account the intercorrelations between items as well as their effect on the reported treatment effect did not favor a one-factor solution. Best fit was achieved through three factors in all three datasets. The factors were similar, but the factorial validity was not perfect either, with some items loading on different factors across datasets. The randomization sequence and the allocation concealment item consistently formed a factor, and the blinding items also consistently formed a second factor across datasets. The other items were not represented by these factors, indicating an additional source of variance. A third factor consistently showed significant loading for acceptable dropout rates. Other items such as original group (intention to treat [ITT]) did also load on this factor but with less consistency across datasets.

The use of checklists when scoring quality versus the application of a summary score has been extensively discussed in the literature. Juni and colleagues (Juni, Witschi, Bloch, et al., 1999; Juni, Douglas, Altman, et al., 2001) raised serious concerns about the use of quantitative sum scores. However, treating all quality items as completely independent does also not appear appropriate either following our analyses. The Cochrane review handbook currently suggests the use of a domain-based evaluation of quality in which critical assessments are made separately for different domains (Higgins and Green, 2009). The equal weighing of each item as applied in our approach is common place but not validated. Depending on the intervention and the clinical field, some internal validity threats may be more pertinent than others; however there are as yet no data to guide what these associations may be. Quality criteria could be used to trigger an overall assessment of quality which is more qualitatively derived than quantitatively by adding individual item scores. The reliability of qualitative overall evaluations have to be considered though, as Hartling et al. (2009) reported a kappa of 0.27 for reviewers to agree on the Cochrane Overall Risk of Bias dimension (Higgins and Green, 2008). A combined qualitative and quantitative approach may be useful: quality features could be ranked by importance for the clinical field a priori and weighted accordingly for a summary score.

Validating scales used to assess the quality of trials is very difficult. The concept of quality is not easy to define and there is no widely accepted gold standard. In one of the datasets we applied the Jadad items and scale and the criteria suggested by Schulz et al. (1995) parallel to our proposed criteria following the original scoring guidelines. We were able to show convergent validity across quality domains. The correlations with the Jadad and the Schulz scales were satisfying throughout and ranged between 0.63 and .0.93. However, the crucial validity test for quality items is the predictive validity of the quality features and possibly scales—is there evidence of bias, and is meeting or not meeting the quality criteria associated with differential effect sizes.

Associations Between Internal Validity and Effect Sizes

This report provides empirical data on the impact of fulfilling or not fulfilling the quality criteria sequence of randomization, concealment of treatment allocation, similarity of baseline values, assessor blinding, care provider blinding, patient blinding, dropout rate, ITT analysis, similarity of cointerventions, acceptable compliance, and similar timing of outcome assessment across three datasets. Although the majority of systematic reviews assess the quality of included studies, and meta-regression analyses trying to trace the effects of quality are often undertaken, there are relatively few published studies showing an effect of quality on effect sizes, that is, empirical evidence of bias in reported study results that can presumably be attributed to the quality of the reporting or the conduct of the research study. For many suggested quality criteria and potential threats to the internal validity of RCTs (see e.g., West et al., 2002; Moja et al., 2005) there is still a dearth of published evidence on the extent of bias, that is, does not meeting the quality criteria show associations with the observed treatment effect.

The 11 proposed quality features contribute information to the evaluation of back pain trials as previously reported (Van Tulder et al., 2009). Although not statistically significant, individual features showed consistently associations with effect sizes depending on the quality of the trial. High-quality studies reported smaller effect sizes, indicating that low-quality studies tended to overestimate treatment effects. A dataset consisting of trials included in EPC reports showed a different pattern. The EPC dataset analysis showed for the majority of individual quality dimensions that the high quality studies in the dataset tended to reported larger treatment effects than the low quality trials that did not meet the quality criterion. The third dataset showed the most variation across quality criteria, but was more similar to the back pain dataset, in that meeting most individual criteria where associated with smaller effect sizes.

The feature allocation concealment showed the most consistent results across datasets. In all three datasets, allocation concealment was associated with effect sizes, and the direction of effect did not vary. Unconcealed trials reported smaller effect sizes in the back pain dataset (effect size difference -0.08, 95% CI: -0.23, 0.07) and the EPC report dataset (effect size difference -0.06, 95% CI: -0.22, 0.11), and the ratio of odds ratios (ROR) in the third dataset also showed that unconcealed trials reported larger treatment effects (ROR=0.91, 95% CI: 0.72, 1.14). This analysis indicated that unconcealed trials tend to overestimate the treatment effect. Similarly, Pidal et al. (2007) reported an ROR of 0.90 (0.81, 1.01); Wood et al. (2008) found a ROR of 0.91 (95% CI 0.80, 1.03) for objective and 0.69 (95% CI: 0.59, 0.82) for subjective outcomes. A pooled analysis using data from Schulz et al. (1995), Moher et al. (1998), Kjaergard et al. (2000), and Juni et al. (2000) showed an ROR of 0.70 (95% CI: 0.62, 0.80) across datasets (Juni et al., 2001). The influence of concealment of allocation on effect size seems to be the most consistent quality criteria.

When applying a total sum score derived from the mean item scores and regressing effect sizes on the sum score, we found no statistically significant linear effect. A simple linear relationship indicates that the reported treatment effects increase the lower the quality level. A similar approach was described by Emerson et al. (1990), who also found no linear relation between quality score and variation in treatment differences.

When using factor scores, rather than individual quality features or a simple sum score, we also did not find that these quality factors predicted effect sizes. The factor structure takes the inter-item correlations as well as their individual association with effect sizes into account. MIMIC models are generally a promising approach to describe complex relations between multiple predictors and has been applied in a variety of fields (e.g., Hartford & Muthén, 2001; Urban & Demetrovics, 2010).

One power maximizing approach showed consistently statistically significant effects. Cutoff values based on the number of fulfilled quality criteria were able to differentiate high- and low-quality studies. In all three datasets, 5 or 6 (out of 11) fulfilled quality criteria differentiated the studies best. In dataset 1 (back pain) and dataset 3 (“pro-bias”), the replication of the set used by Moher et al. (1998) high-quality studies showed smaller treatment effects. Effect size differences were -0.20 in dataset #1 and the RORs were 0.79 (cutoff at 5) and 0.77 (cutoff at 6) in the third dataset when taking several quality criteria into account. However, in the EPC report dataset, we found unexpected results: studies with five or six fulfilled quality feature reported larger effect sizes than the low-quality trials (effect size difference 0.27). Moher et al. (1998) used the Jadad scale, which takes three individual quality features into account (randomization, blinding, withdrawals) to differentiate high- and low-quality studies, and reported a ratio of odds ratio of 0.66 (95% CI: 0.52, 0.83). Our results are similar in direction for two of the three datasets, but we found less-pronounced results. The difference in results may be due to the selected statistical approach and presumably in part due to the nonperfect overlap of studies since we were unable to obtain a quarter of the original dataset.

Comparing results across methods to test associations between quality criteria and effect sizes, the differences between the originally analyzed back pain dataset and the EPC report dataset are most striking. The quality criteria were developed for the CBRG and they also appear to be most useful in this dataset (see also van Tulder et al., 2003). However, it is possible that the tendency of EPC reports to consider quality as an inclusion criterion, potentially excluding fatally flawed studies from the review, is partly responsible for the different results in this dataset. The quality scores indicated that for several criteria such as blinding, similarity of cointerventions, and ITT analysis, the trials included in EPC reports scored higher.

Moderators and Confounders

The identified associations between quality and effect sizes varied across our datasets presented in this report which reflects also the extent of conflicting results reported in the literature (e.g., Moher et al., 1998, Balk et al., 2002; Juni et al., 2001). Research on the association between quality and effect sizes should focus on factors that can help predict when lack of quality is likely to result in a distorted estimate of treatment effects.

In this report, we investigated a number of differences across datasets. The systematically investigated moderators and confounders were the size of the treatment effect within meta-analyses, the condition being treated, the type of outcome, and the variance in effect sizes. These moderators did not sufficiently explain diverging results across our employed datasets.

The condition being treated or the clinical field the study was conducted in was not sufficient to explain differential effects of the association between quality and effect sizes. The size of the treatment effect could also not be shown as a significant moderator between the two variables. Unlike Wood et al. (2009), we could not show that the type of outcome explained differences in effects of associations; there was no statistically significant difference between slopes. However, it has to be noted that the type of outcome is to some extent already been taking into account in the CBRG guidance: assessor blinding is assumed for studies with automated test result analysis, it is assumed that the assessor is clearly not aware of the treatment allocation, regardless of whether the publication states that the assessor was blind.

The variance in effect sizes across a dataset was one factor that should be explored further in future research to see if this factor contributes to the question of when quality features are likely to influence effect sizes. Balk et al. (2002) used existing heterogeneity in odds ratios as an inclusion criterion for their meta-epidemiological study but concluded that the investigated quality measure are not reliable associated with the strength of treatment effect. Whether the variance in effect sizes is indeed a sufficient moderator to explain variation across datasets is a testable question and could be assessed with Monte Carlo simulations systematically investigating the effect of moderators or confounders that influence the association between quality and effect sizes.

Implication for Practice

In two of these last datasets, we showed that quality features can affect reported treatment effects. The 11 proposed quality features developed for the CBRG contribute information to the evaluation of back pain trials as previously reported (Van Tulder et al., 2009). Whether this extended list of quality features can be proposed for a more general use was one of the principal questions of this research project. Their general applicability has not been supported, as their effect was not uniform across datasets.

We conclude from our analyses that the association between quality features and reported treatment effects should be explored in systematic reviews. Regardless of whether quality criteria are assessed individually, through empirically derived factor scores or the use of a total score, regardless of whether the summary score was quantitatively, quantitative or through a combination derived, and regardless of whether low quality studies are excluded from the analysis or studies are pooled weighted by quality (e.g. Juni, Altman, and Egger, 2001; Welton, Ades, Altman, et al., 2009) quality should be taken into account when evaluating the existing evidence and the potential bias should be quantified. For situations where the Jadad criteria may be insufficiently applicable, our data provide some support for the use of the II-item CBRG list.

Future Research

Applying psychometric principles to the field of quality criteria is rarely explored but can provide useful insight into empirical associations of quality items. In this report we explored the reliability of the proposed quality items only through item and scale analysis. Future work should include an analysis of agreement between raters. The reproducibility of quality judgments across independent raters is another valuable method for estimating the reliability of proposed items or scales. Previous research has shown that also carefully developed tools may show disappointing rater agreement when scoring agreement is tested. Hartling et al. (2009) reported kappas ranging from 0.13 to 0.74 for domains of the Cochrane Risk of Bias tool. For evidence reviews, however, we suggest an additional approach for testing the reliability of tools. In systematic reviews, it is now standard to employ two independent raters when scoring the quality of included studies and to reconcile independent decisions for a final score. This approach helps to avoid individual reviewer bias and errors and the reconciled decision should be more reliable than the individual decision. The rater agreement of reconciled decisions across pairs of raters is a better indicator of the reliability of the tool because it mirrors more closely how the tool will be used in practice.

There is a need for more information on individual quality features and empirical evidence of bias. This concerns the many suggested quality criteria for which no empirical evidence is available yet or at least no summary across individual meta-analyses exists (see West et al., 2002; Moja et al., 2005). There are other quality criteria such as selective outcome reporting

(Kirkham, Dwan, Altman, et al., 2010) that may be difficult to operationalize and a replication of the effect in a different dataset would be useful. The scarcity of evidence is disappointing considering the fact that quality scoring is a standard method in systematic reviews. Many different quality criteria are used by reviewers and many represent possible and plausible threats to the validity of the study. Future reviews should report data on these associations in order to advance the evidence base for quality assessments.

The association between quality features and effect sizes is complex, and the conditions, when lack of quality is most likely to lead to bias, should be explored further in future research. Factors such as the variance in quality scores and effect sizes across studies could be systematically studied in “virtual datasets,”, that is, by creating datasets employing Monte Carlo simulation methods. Using datasets of “known” properties would be useful to further study associations between the proposed quality criteria and effect sizes. An increase in sample size and thereby statistical power would enable researchers to detect small but systematic effects and shed further light on the question of when quality features are most useful to be taken into account when assessing treatment effects in published research.

Conclusions

The associations between quality features and effect sizes are complex. Effect sizes of individual studies depend on many factors. In two datasets, individual quality items and summary scores of items were associated with differences in effect sizes. This relationship was not found in the remaining dataset. Despite several exploratory analyses, we were not able to explain these differences. The conditions under which quality features and which features lead to biased effect sizes warrant further exploration, and factors such as the variance in quality scores and effect sizes will be investigated in a subsequent project.

References

Assendelft WJ, Morton SC, Yu EI, et al. Spinal manipulative therapy for low back pain. A meta-analysis of effectiveness relative to other therapies. Ann Intern Med 2003 Jun 3;138(11):871–81.

Balk E, Tatsioni A, Lichtenstein A, et al. Effect of chromium supplementation on glucose metabolism and lipids: a systematic review of randomized controlled trials. Diabetes Care 2007 Aug;30(8):2154–63.

Balk EM, Bonis PA, Moskowitz H, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002 Jun 12;287(22):2973–82.

Balk EM, Lichtenstein AH, Chung M, et al. Effects of omega-3 fatty acids on serum markers of cardiovascular disease risk: a systematic review. Atherosclerosis 2006 Nov;189(1):19–30.

Berkey CS, Hoaglin DC, Mosteller F, et al. A random-effects regression model for meta-analysis. Stat Med 1995 Feb 28;14(4):395–411.

Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995 Jan 21;310(6973):170.

Chapell R, Reston J, Snyder D. Management of Treatment-Resistant Epilepsy. Evidence Report/Technology Assessment No. 77 (Prepared by ECRI Evidence-based Practice Center under Contract No. 290-97-0020). Rockville, MD: Agency for Healthcare Research and Quality, May 2003. AHRQ Publication No. 03-0028.

Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. I: Medical. Stat Med 1989 Apr;8(4):441–54.

Coulter I, Hardy M, Shekelle P, al. Effect of the Supplemental Use of Antioxidants Vitamin C, Vitamin E, and Coenzyme Q10 for the Prevention and Treatment of Cancer. Evidence Report/ Technology Assessment No. 75 (Prepared by Southern California Evidence-based Practice Center under Contract No. 290-97-0001). Rockville, MD: Agency for Healthcare Research and Quality, August 2003. AHRQ Publication No. 03-E047.

Counsell C, Sandercock P. Use of anticoagulants in patients with acute ischemic stroke. Stroke 1995 Mar;26(3):522–3.

Donahue K, Gartlehner G, Jonas D, et al. Comparative Effectiveness of Drug Therapy for Rheumatoid Arthritis and Psoriatic Arthritis in Adults. Comparative Effectiveness Review No. 11 (Prepared by RTI-University of North Carolina Evidence-based Practice Center under Contract No. 290-02-0016). Rockville, MD: Agency for Healthcare Research and Quality, November 2007. AHRQ Publication No. 08-EHC004-EF.

Egger M, Juni P, Bartlett C, et al. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7(1):1–76.

Emerson JD, Burdick E, Hoaglin DC, et al. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Control Clin Trials 1990 Oct;11(5):339–52.

Furlan AD, Imamura M, Dryden T, et al. Massage for low-back pain. Cochrane Database Syst Rev. 2008(4):CD001929.

Furlan AD, van Tulder M, Cherkin D, et al. Acupuncture and dry-needling for low back pain: an updated systematic review within the framework of the cochrane collaboration. Spine (Phila Pa 1976) 2005 Apr 15;30(8):944–63.

Hagen KB, Hilde G, Jamtvedt G, et al. Bed rest for acute low back pain and sciatica. Nurs Times 2001 Aug 2–8;97(31):40.

Hansen RA, Gartlehner G, Webb AP, et al. Efficacy and safety of donepezil, galantamine, and rivastigmine for the treatment of Alzheimer's disease: a systematic review and meta-analysis. Clin Interv Aging 2008;3(2):211–25.

Hardy M, Coulter I, Morton SC, et al. S-Adenosyl-L-Methionine for Treatment of Depression, Osteoarthritis, and Liver Disease. Evidence Report/Technology Assessment No. 64 (Prepared by the Southern California Evidence-based Practice Center under Contract No. 290-97-0001). Rockville, MD: Agency for Healthcare Research and Quality, October 2002. AHRQ Publication No. 02-E034.

Harford TC, Muthen BO. The dimensionality of alcohol abuse and dependence: a multivariate analysis of DSM-IV symptom items in the National Longitudinal Survey of Youth. J Stud Alcohol 2001 Mar;62(2):150–7.

Hartling L, Ospina M, Liang Y, et al. Risk of bias versus quality assessment of randomised controlled trials: cross sectional study. BMJ 2009;339:b4012.

Hayden JA, van Tulder MW, Malmivaara A, et al. Exercise therapy for treatment of non-specific low back pain. Cochrane Database Syst Rev 2005(3):CD000335.

Henschke N, Ostelo RW, van Tulder MW, et al. Behavioural treatment for chronic low-back pain. Cochrane Database Syst Rev 2010;7:CD002014.

Heymans MW, van Tulder MW, Esmail R, et al. Back schools for nonspecific low back pain: a systematic review within the framework of the Cochrane Collaboration Back Review Group. Spine (Phila Pa 1976) 2005 Oct 1;30(19):2153–63.

Higgins J, Green S. Cochrane handbook for systematic reviews of interventions version 5.0.2. [updated September 2009]. Cochrane Collaboration. 2008.

Hughes E, Collins J, P. V. Bromocriptine for unexplained subfertility in women. Cochrane Database Syst Rev 1996;4.

Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996 Feb;17(1):1–12.

Juni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. BMJ 2001 Jul 7;323(7303):42–6.

Juni P, Tallon D, Egger M. Garbage in - garbage out? Assessment of the quality of controlled trials in meta-analyses published in leading journals. Proceedings of the 3rd symposium on systematic reviews: beyond the basics, St Catherine's College, Oxford Oxford: Centre for Statistics in Medicine. 2000:19.

Juni P, Witschi A, Bloch R, et al. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999 Sep 15;282(11):1054–60.

Kane RL, Wang J, Garrard J. Reporting in randomized clinical trials improved after adoption of the CONSORT statement. J Clin Epidemiol 2007 Mar;60(3):241–9.

Karjalainen K, Malmivaara A, van Tulder M, et al. Multidisciplinary biopsychosocial rehabilitation for subacute low back pain in working-age adults: a systematic review within the framework of the Cochrane Collaboration Back Review Group. Spine (Phila Pa 1976) 2001 Feb 1;26(3):262–9.

Khadilkar A, Milne S, Brosseau L, et al. Transcutaneous electrical nerve stimulation (TENS) for chronic low-back pain. Cochrane Database Syst Rev 2005(3):CD003008.

Kirkham JJ, Dwan KM, Altman DG, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ 2010;340:c365.

Kjaergard LL, Villumsen J, Gluud C. Quality of randomised clinical trials affects estimates of intervention efficacy. Proceedings of the 7th Cochrane colloquium Universita STommaso D'Aquino, Rome Milan: Centro Cochrane Italiano 1999:p. 57 (poster B10).

Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 2001 Dec 4;135(11):982–9.

Lensing AW, Prins MH, Davidson BL, et al. Treatment of deep venous thrombosis with low-molecular-weight heparins. A meta-analysis. Arch Intern Med 1995 Mar 27;155(6):601–7.

Lo GH, LaValley M, McAlindon T, et al. Intra-articular hyaluronic acid in treatment of knee osteoarthritis: a meta-analysis. JAMA 2003 Dec 17;290(23):3115–21.

Loonen AJ, Peer PG, Zwanikken GJ. Continuation and maintenance therapy with antidepressive agents. Meta-analysis of research. Pharm Weekbl Sci 1991 Aug 23;13(4):167–75.

Loosemore TM, Chalmers TC, Dormandy JA. A meta-analysis of randomized placebo control trials in Fontaine stages III and IV peripheral occlusive arterial disease. Int Angiol 1994 Jun;13(2):133–42.

Mari JJ, Streiner DL. An overview of family interventions and relapse on schizophrenia: meta-analysis of research findings. Psychol Med 1994 Aug;24(3):565–78.

Marshall JK, Irvine EJ. Rectal aminosalicylate therapy for distal ulcerative colitis: a meta-analysis. Aliment Pharmacol Ther 1995 Jun;9(3):293–300.

Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998 Aug 22;352(9128):609–13.

Moja LP, Telaro E, D'Amico R, et al. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ 2005 May 7;330(7499):1053.

Pace F, Maconi G, Molteni P, et al. Meta-analysis of the effect of placebo on the outcome of medically treated reflux esophagitis. Scand J Gastroenterol 1995 Feb;30(2):101–5.

Pildal J, Hrobjartsson A, Jorgensen KJ. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. Int J Epidemiol 2007 Aug;36(4):847–57.

Ramirez-Lassepas M, Cipolle RJ. Medical treatment of transient ischemic attacks: does it influence mortality? Stroke 1988 Mar;19(3):397–400.

Roelofs PD, Deyo RA, Koes BW. Nonsteroidal anti-inflammatory drugs for low back pain: an updated Cochrane review. Spine (Phila Pa 1976) 2008 Jul 15;33(16):1766–74.

Schulz KF, Chalmers I, Hayes RJ, et al. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995 Feb 1;273(5):408–12.

Shekelle P, Hardy ML, Coulter I, et al. Effect of the supplemental use of antioxidants vitamin C, vitamin E, and coenzyme Q10 for the prevention and treatment of cancer. Evid Rep Technol Assess (Summ) 2003 Oct(75):1–3.

Shekelle PG, Maglione M, Bagley S, et al. Efficacy and Comparative Effectiveness of Off-Label Use of Atypical Antipsychotics. Comparative Effectiveness Review No. 6 (Prepared by the Southern California/RAND Evidence-based Practice Center under Contract No. 290-02-0003). Rockville, MD: Agency for Healthcare Research and Quality, January 2007. AHRQ Publication No.07-EHC003-EF.

Shekelle PG, Morton SC, Maglione M, et al. Pharmacological and Surgical Treatment of Obesity. Evid Report/Technology Assessment No. 103 (Prepared by the Southern California/RAND Evidence-based Practice Center under Contract No. 290-02-0003). Rockville, MD: Agency for Healthcare Research and Quality, July 2004. AHRQ Publication No. 04-E028-2.

Sutherland LR, May GR, Shaffer EA. Sulfasalazine revisited: a meta-analysis of 5-aminosalicylic acid in the treatment of ulcerative colitis. Ann Intern Med 1993 Apr 1;118(7):540–9.

Towfigh A, Romanova M, Weinreb JE, et al. Self-monitoring of blood glucose levels in patients with type 2 diabetes mellitus not taking insulin: a meta-analysis. Am J Manag Care 2008 Jul;14(7):468–75.

Urban R, Demetrovics Z. Smoking outcome expectancies: A multiple indicator and multiple cause (MIMIC) model. Addict Behav 2010 Jun;35(6):632–5.

Van Duijvenbode IC, Jellema P, van Poppel MN, et al. Lumbar supports for prevention and treatment of low back pain. Cochrane Database Syst Rev 2008(2):CD001823.

Van Tulder M, Furlan A, Bombardier C, et al. Updated method guidelines for systematic reviews in the cochrane collaboration back review group. Spine (Phila Pa 1976) 2003 Jun 15;28(12):1290–9.

Van Tulder MW, Suttorp M, Morton S, et al. Empirical evidence of an association between internal validity and effect size in randomized controlled trials of low-back pain. Spine (Phila Pa 1976) 2009 Jul 15;34(16):1685–92.

Van Tulder MW, Touray T, Furlan AD, et al. Muscle relaxants for nonspecific low back pain: a systematic review within the framework of the cochrane collaboration. Spine (Phila Pa 1976) 2003 Sep 1;28(17):1978–92.

Verhagen AP, de Vet HC, de Bie RA, et al. The Delphi list: a criteria list for quality assessment of randomized clinical trials for conducting systematic reviews developed by Delphi consensus. J Clin Epidemiol 1998 Dec;51(12):1235–41.

Welton N, Ades A, Carlin J, et al. Models for potentially biased evidence in meta-analysis using empirically based priors. Journal of the Royal Statistical Society: Series A (Statistics in Society) 2009;172:119–36.
West S, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. Evid Rep Technol Assess (Summ) 2002 Mar(47):1–11.

Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008 Mar 15;336(7644):601–5.

Appendix A. References Dataset 1: Back Pain, 216 Trials

Agrifoglio E, Benvenuti M, Gatto P, et al. Aceclofenac: A New NSAID in the Treatment of Acute Lumbago Multicentre Single Blind Study vs. Diclofenac. Acta Therapeutica 1994;20:33–41.

Alexandre NM, de Moraes MA, Correa Filho, HR, et al. Evaluation of a program to reduce back pain in nursing personnel. Rev Saude Publica 2001;35:356–61.

Altmaier EM, Lehmann TR, Russell DW, et al. The effectiveness of psychological interventions for the rehabilitation of low back pain: a randomized controlled trial evaluation. Pain 1992;49:329–35.

Amlie E, Weber H, Holme I. Treatment of acute low-back pain with piroxicam: results of a double-blind placebo-controlled trial. Spine (Phila Pa 1976) 1987;12:473–6.

Andersson GB, Lucente T, Davis AM, et al. A comparison of osteopathic spinal manipulation with standard care for patients with low back pain. N Engl J Med 1999;341:1426–31.

Aoki T, Kuroki Y, Kageyama T, et al. Multicentre double-blind comparison of piroxicam and indomethacin in the treatment of lumbar diseases. Eur J Rheumatol Inflamm 1983;6:247–52.

Arbus L, Fajadet B, Aubert D, et al. Activity of Tetrazepam (Myolastan) in Low Back Pain. A Double-Blind Trial v. Placebo. Clinical Trials Journal 1990;27:258–67.

Aure OF, Nilsen JH, Vasseljen O. Manual therapy and exercise therapy in patients with chronic low back pain: a randomized, controlled trial with 1-year follow-up. Spine (Phila Pa 1976) 2003;28:525–31; discussion 531–2.

Babej-Dolle R, Freytag S, Eckmeyer J, et al. Parenteral dipyrone versus diclofenac and placebo in patients with acute lumbago or sciatic pain: randomized observer-blind multicenter study. Int J Clin Pharmacol Ther 1994;32:204–9.

Bakshi R, Broll H, Klein G, et al. Treatment of Acute Lumbosacral Back Pain with Diclofenac Resinate. Drug Invest 1994;8:288–93.

Baratta RR. A double-blind study of cyclobenzaprine and placebo in the treatment of acute musculoskeletal conditions of the low back. Current Therapeutic Research 1982;32:646–52.

Basler HD, Jakle C, Kroner-Herwig B. Incorporation of cognitive-behavioral treatment into the medical care of chronic low back patients: a controlled randomized study in German pain treatment centers. Patient Educ Couns 1997;31:113–24.

Basmajian JV. Acute back pain and spasm. A controlled multicenter trial of combined analgesic and antispasm agents. Spine (Phila Pa 1976) 1989;14: 438–9.

Bendix AF, Bendix T, Lund C, et al. Comparison of three intensive programs for chronic low back pain patients: a prospective, randomized, observer-blinded study with one-year follow-up. Scand J Rehabil Med 1997;29:81–9.

Bendix AF, Bendix T, Ostenfeld S, et al. Active treatment programs for patients with chronic low back pain: a prospective, randomized, observer-blinded study. Eur Spine J 1995;4:148–52.

Bendix AF, Bendix T, Vaegter K, et al. Multidisciplinary intensive treatment for chronic low back pain: a randomized, prospective study. Cleve Clin J Med 1996;63:62–9.

Bendix T, Bendix A, Labriola M, et al. Functional restoration versus outpatient physical training in chronic low back pain: a randomized comparative study. Spine (Phila Pa 1976) 2000;25:2494–2500.

Bentsen H, Lindgarde F, Manthorpe R. The effect of dynamic strength back exercise and/or a home training program in 57-year-old women with chronic low back pain. Results of a prospective randomized study with a 3-year follow-up period. Spine (Phila Pa 1976) 1997;22: 1494–1500.

Bergquist-Ullman M, Larsson U. Acute low back pain in industry. A controlled prospective study with special reference to therapy and confounding factors. Acta Orthop Scand 1977;1–117.

Berry H, Hutchinson DR. A multicentre placebo-controlled study in general practice to evaluate the efficacy and safety of tizanidine in acute low-back pain. J Int Med Res 1988;16:75–82.

Berry H, Hutchinson DR. Tizanidine and ibuprofen in acute low-back pain: results of a double-blind multicentre study in general practice. J Int Med Res 1988;16:83–91.

Berwick DM, Budman S, Feldstein M. No Clinical Effect of Back Schools in an HMO. A Randomized Prospective Trial. Spine 1989;14.

Bianchi M. Evaluation of Cyclobenzaprine for skeletal muscle spasm of local origin. In: Clinical Evaluation of Flexeril (Cyclobenzaprine HCl/MSD). Minneapolis, MN: Postgraduate Medicine Communications; 1978. P. 25–9.

Blazek M, Keszthelyi B, Varhelyi M, et al. Comparative study of Biarison and Voltaren in acute lumbar pain and lumbo-ischialgia. Ther Hung 1986;34:163–6.

Blomberg S, Hallin G, Grann K, et al. Manual therapy with steroid injections--a new approach to treatment of low back pain. A controlled multicenter trial with an evaluation by orthopedic surgeons. Spine (Phila Pa 1976) 1994;19:569–77.

Borenstein DG, Lacks S, Wiesel SW. Cyclobenzaprine and naproxen versus naproxen alone in the treatment of acute low back pain and muscle spasm. Clin Ther 1990;12:125–31.

Boyles WF, Glassman JM, Soyka JP. Management of Acute Musculoskeletal Conditions: Thoracolumbar Strain or Sprain: A Double-Blind Evaluation Comparing the Efficacy and Safety of Carisoprodol with Diazepam. Today's Ther. Trends 1983;1–16.

Braun H and Huberty R. [Therapy of lumbar sciatica. A comparative clinical study of a corticoid-free monosubstance and a corticoid-containing combination drug]. Med Welt 1982;33:490–1.

Bronfort G. Chiropractic versus general medical treatment of low back pain: A small scale controlled clinical trial. AJCM 1989;2:145–50.

Bronfort G, Goldsmith CH, Nelson CF, et al. Trunk exercise combined with spinal manipulative or NSAID therapy for chronic low back pain: a randomized, observer-blinded clinical trial. J Manipulative Physiol Ther 1996;19:570–82.

Brown FL Jr, Bodison S, Dixon J, et al. Comparison of diflunisal and acetaminophen with codeine in the treatment of initial or recurrent acute low back strain. Clin Ther 1986;9 Suppl C:52–8.

Bru E, Mykletun RJ, Svebak S. (1994) Assessment of musculoskeletal and other health complaints in female hospital staff. Appl Ergon 1994;25:101–5.

Bruggemann G, Koehler CO, Koch EM. [Results of a double-blind study of diclofenac + vitamin B1, B6, B12 versus diclofenac in patients with acute pain of the lumbar vertebrae. A multicenter study]. Klin Wochenschr 1990;68:116–20.

Buswell J. Low Back Pain: a comparison of two treatment programmes. NZ Journal of Physiotherapy 1982;13-17.

Calmels P, Jacob JF, Fayolle-Minon I, et al. [Use of isokinetic techniques vs standard physiotherapy in patients with chronic low back pain. Preliminary results]. Ann Readapt Med Phys 2004;47:20–7.

Carlsson CP and Sjolund BH. Acupuncture for chronic low back pain: a randomized placebo-controlled study with long-term follow-up. Clin J Pain 2001;17:296–305.

Ceccherelli F, Rigoni MT, Gagliardi G et al. Comparison of superficial and deep acupuncture in the treatment of lumbar myofascial pain: a double-blind randomized controlled study. Clin J Pain 2002;18:149–53.

Cherkin DC, Deyo RA, Battie M, et al. A comparison of physical therapy, chiropractic manipulation, and provision of an educational booklet for the treatment of patients with low back pain. N Engl J Med 1998;339:1021–9.

Cherkin DC, Eisenberg D, Sherman KJ, et al. Randomized trial comparing traditional Chinese medical acupuncture, therapeutic massage, and self-care education for chronic low back pain. Arch Intern Med 2001;161:1081–8.

Chok B, Lee R, Latimer J, et al. Endurance training of the trunk extensor muscles in people with subacute low back pain. Phys Ther 1999;79:1032–42.

Coan RM, Wong G, Ku SL, et al. The acupuncture treatment of low back pain: a randomized controlled study. Am J Chin Med 1980;8:181–9.

Colberg K, Hettich M, Sigmund R et al. The efficacy and tolerability of an 8-day administration of intravenous and oral meloxicam: a comparison with intramuscular and oral diclofenac in patients with acute lumbago. German Meloxicam Ampoule Study Group. Curr Med Res Opin 1996;13:363–77.

Coomes EN. A comparison between epidural anaesthesia and bed rest in sciatica. Br Med J 1961;1:20–4.

Corey DT, Koepfler LE, Etlin D, et al. A limited functional restoration program for injured workers: A randomized trial. Journal of Occupational Rehabilitation 1996;6:239–49.

Corts Giner JR. Estudio DS 103–282: Relajante Muscular en Lumbalgia Aguda o Lumbago (Estudio Doble Ciego de Tizanidina plus Paracetamol Vs Placebo plus Paracetamol). Rev. Esp. de Cir. Ost 1989;119–24.

Dalichau S, Scheele K. [Effects of elastic lumbar belts on the effect of a muscle training program for patients with chronic back pain]. Z Orthop Ihre Grenzgeb 2000;138:8–16.

Dalichau S, Scheele K, Perry RM, et al. Ultrasound-aided posture-and motion analysis of the lumbar spine for the evaluation of the efficacy of a back school in the construction industry. Zbl Arbeitsmed 1999;49:148–56.

Dapas F, Hartman SF, Martinez L, et al. Baclofen for the treatment of acute low-back syndrome. A double-blind comparison with placebo. Spine (Phila Pa 1976) 1985;10:345–9.

Davies JE, Gibson T, Tester L. The value of exercises in the treatment of low back pain. Rheumatol Rehabil 1979;18:243–7.

Davoli L, Ciotti G, Biondi M, et al. Piroxicam-Beta-Cyclodextrin in the Treatment of Low-Back-Pain. Current Therapeutic Research 1989;46:940–7.

Delitto A, Cibulka MT, Erhard RE, et al. Evidence for use of an extension-mobilization category in acute low back syndrome: a prescriptive validation pilot study. Phys Ther 1993;73:216–22; discussion 223–8.

Descarreaux M, Normand MC, Laurencelle L, et al. Evaluation of a specific home exercise program for low back pain. J Manipulative Physiol Ther 2002;25:497–503.

Dettori JR, Bullock SH, Sutlive TG, et al. The effects of spinal flexion and extension exercises and their associated postures in patients with acute low back pain. Spine (Phila Pa 1976) 1995;20:2303–12.

Deyo RA, Diehl AK, Rosenthal M. How many days of bed rest for acute low back pain? A randomized clinical trial. N Engl J Med 1986;315:1064–70.

Deyo RA, Walsh NE, Martin DC, et al. A controlled trial of transcutaneous electrical nerve stimulation (TENS) and exercise for chronic low back pain. N Engl J Med 1990;322:1627–34.

Donaldson S, Romney D, Donaldson M, et al. Randomized study of the application of single motor unit biofeedback training to chronic low pain. Journal of Occupational Rehabilitation 1994;4:23–37.

Doran DM, Newell DJ. Manipulation in treatment of low back pain: a multicentre study. Br Med J 1975;2:161–4.

Driessens M, Famaey JP, Orloff S, et al. Efficacy and tolerability of sustained-release ibuprofen in the treatment of patients with chronic back pain. Current Therapeutic Research 1994;55:1283–92.

Edelist G, Gross AE and Langer F. Treatment of low back pain with acupuncture. Can Anaesth Soc J 1976;23:303–6.

Elnaggar IM, Nordin M, Sheikhzadeh A, et al. Effects of spinal flexion and extension exercises on low-back pain and spinal mobility in chronic mechanical low-back pain patients. Spine (Phila Pa 1976) 1991;16:967–72.

Erhard RE, Delitto A, Cibulka MT. Relative effectiveness of an extension program and a combined program of manipulation and flexion and extension exercises in patients with acute low back syndrome. Phys Ther 1994;74:1093–1100.

Evans DP, Burke MS, Lloyd KN, et al. Lumbar spinal manipulation on trial. Part I--clinical assessment. Rheumatol Rehabil 1978;17:46–53.

Farrell JP, Twomey LT. Acute low back pain. Comparison of two conservative treatment approaches. Med J Aust 1982;1:160–4.

Franke A, Gebauer S, Franke K, et al. [Acupuncture massage vs Swedish massage and individual exercise vs group exercise in low back pain sufferers--a randomized controlled clinical trial in a 2 x 2 factorial design]. Forsch Komplementarmed Klass Naturheilkd 2000;7:286–93.

Friedrich M, Gittler G, Halberstadt Y, et al. Combined exercise and motivation program: effect on the compliance and level of disability of patients with chronic low back pain: a randomized controlled trial. Arch Phys Med Rehabil 1998;79:475–87.

Frost H, Klaber Moffett JA, Moser JS, et al. Randomised controlled trial for evaluation of fitness programme for patients with chronic low back pain. BMJ 1995;310:151–4.

Galantino ML, Bzdewka TM, Eissler-Russo JL, et al. The impact of modified Hatha yoga on chronic low back pain: a pilot study. Altern Ther Health Med 2004;10:56–9.

Garvey TA, Marks MR, Wiesel SW. A prospective, randomized, double-blind evaluation of trigger-point injection therapy for low-back pain. Spine (Phila Pa 1976) 1989;14:962–4.

Gemignani G, Olivieri I, Ruju G, et al. Transcutaneous electrical nerve stimulation in ankylosing spondylitis: a double-blind study. Arthritis Rheum 1991;34:788–9.

Gibson T, Grahame R, Harkness J, et al. Controlled comparison of short-wave diathermy treatment with osteopathic treatment in non-specific low back pain. Lancet 1985;1:1258–61.

Gilbert JR, Taylor DW, Hildebrand A et al. Clinical trial of common treatments for low back pain in family practice. Br Med J (Clin Res Ed) 1985;291:791–4.

Giles LG, Muller R. Chronic spinal pain: a randomized clinical trial comparing medication, acupuncture, and spinal manipulation. Spine (Phila Pa 1976) 2003;28:1490–1502; discussion 1502–3.

Giles LG, Muller R. Chronic spinal pain syndromes: a clinical pilot trial comparing acupuncture, a nonsteroidal anti-inflammatory drug, and spinal manipulation. J Manipulative Physiol Ther 1999;22:376–81.

Glassman J. Management of acute musculoskeletal conditions-thoracolumbar strain or sprain: A double-blind evaluation comparing the efficacy and safety of carisoprodol with cyclobenzaprine hydrochloride. Current Therapeutic Research 1983;34:917–29.

Godfrey CM, Morgan PP, Schatzker J. A randomized trial of manipulation for low-back pain in a medical setting. Spine (Phila Pa 1976) 1984;9:301–4.

Gold R. Orphenadrine Citrate: Sedative or Muscle Relaxant. Clinical Therapeutics 1978;1:451–3.

Goldie I. A clinical trial with indomethacin (indomee(R)) in low back pain and sciatica. Acta Orthop Scand 1968;39:117–28.

Grant DJ, Bishop-Miller J, Winchester DM, et al. A randomized comparative trial of acupuncture versus transcutaneous electrical nerve stimulation for chronic back pain in the elderly. Pain 1999;82:9–13.

Gunn CC, Milbrandt WE, Little AS et al. Dry needling of muscle motor points for chronic low-back pain: a randomized clinical trial with long-term follow-up. Spine (Phila Pa 1976) 1980;5:279–91.

Gur A, Karakoc M, Cevik R, et al. Efficacy of low power laser therapy and exercise on pain and functions in chronic low back pain. Lasers Surg Med 2003;32:233–8.

Hadler NM, Curtis P, Gillings DB, et al. A benefit of spinal manipulation as adjunctive therapy for acute low-back pain: a stratified controlled trial. Spine (Phila Pa 1976) 1987;12:702–6.

Hansen FR, Bendix T, Skov P, et al. Intensive, dynamic back-muscle exercises, conventional physiotherapy, or placebo-control treatment of low-back pain. A randomized, observer-blind trial. Spine (Phila Pa 1976) 1993;18:98–108.

Harkapaa K, Jarvikoski A, Mellin G, et al. A controlled study on the outcome of inpatient and outpatient treatment of low back pain. Part I. Pain, disability, compliance, and reported treatment benefits three months after treatment. Scand J Rehabil Med 1989;21:81–9.

Hemmila HM, Keinanen-Kiukaanniemi SM, Levoska S, et al. Does folk medicine work? A randomized clinical trial on patients with prolonged back pain. Arch Phys Med Rehabil 1997;78:571–7.

Hennies OL. A new skeletal muscle relaxant (DS 103-282) compared to diazepam in the treatment of muscle spasm of local origin. J Int Med Res 1981;9:62–8.

Hernandez-Reif M, Field T, Krasnegor J et al. Lower back pain is reduced and range of motion increased after massage therapy. Int J Neurosci 2001;106:131–45.

Herzog W, Conway PJ, Willcox BJ. Effects of different treatment modalities on gait symmetry and clinical measures for sacroiliac joint patients. J Manipulative Physiol Ther 1991;14:104–9.

Hickey RF. Chronic low back pain: a comparison of diflunisal with paracetamol. N Z Med J 1982;95:312–4.

Hides JA, Jull GA, Richardson CA. Long-term effects of specific stabilizing exercises for first-episode low back pain. Spine (Phila Pa 1976) 2001;26:E243–8.

Hildebrandt VH, Proper KI, van den Berg R, et al. [Cesar therapy is temporarily more effective in patients with chronic low back pain than the standard treatment by family practitioner: randomized, controlled and blinded clinical trial with 1 year follow-up]. Ned Tijdschr Geneeskd 2000;144:2258–64.

Hindle TH 3rd. Comparison of carisoprodol, butabarbital, and placebo in treatment of the low back syndrome. Calif Med 1972;117:7–11.

Hingorani K. Diazepam in backache. A double-blind controlled trial. Ann Phys Med 1966;8:303–6.

Hingorani K. Orphenadrin-paracetamol in backache-a double-blind controlled trial. Br J Clin Pract 1971;25:227–31.

Hingorani K, Biswas AK. Double-blind controlled trial comparing oxyphenbutazone and indomethacin in the treatment of acute low back pain. Br J Clin Pract 1970;24:120–3.

Hoehler FK, Tobis JS, Buerger AA. Spinal manipulation for low back pain. JAMA 1981;245:1835–8.

Hosie G. The topical NSAID, felbinac, versus oral ibuprofen: a comparison of efficacy in the treatment of acute lower back injury. British Journal of Clinical Research 1993;4:5–17.

Hsieh CY, Adams AH, Tobis J, et al. Effectiveness of four conservative treatments for subacute low back pain: a randomized clinical trial. Spine (Phila Pa 1976) 2002;27:1142–8.

Hume Kendall P, Jenkins JM. Exercises for Backache: A Double-Blind Controlled Trial. Physical Medicine Department, Guy's Hospital London. Physiotherapy 1968;54:154–7.

Hurri H. The Swedish back school in chronic low back pain. Part I. Benefits. Scand J Rehabil Med 1989;21:33–40.

Indahl A, Velund L, Reikeraas O. Good prognosis for low back pain when left untampered. A randomized clinical trial. Spine (Phila Pa 1976) 1995;20:473–7.

Jaffe G. A double-blind, between-patient comparison of alclofenac ('Prinalgin') and indomethacin in the treatment of low back pain and sciatica. Curr Med Res Opin 1974;2:424–9.

Transcutaneous Electrical Nerve Stimulation (TENS) For Chronic Low Back Pain. Annual meeting of the American Academy of Orthopedic Surgeons; 1997; San Francisco.

Johannsen F, Remvig L, Kryger P, et al. Exercises for chronic low back pain: a clinical trial. J Orthop Sports Phys Ther 1995;22:52–9.

Jousset N, Fanello S, Bontoux L, et al. Effects of functional restoration versus 3 hours per week physical therapy: a randomized controlled study. Spine (Phila Pa 1976) 2004;29: 487–93; discussion 494.

Kankaanpaa M, Taimela S, Airaksinen O, et al. The efficacy of active rehabilitation in chronic low back pain. Effect on pain intensity, self-experienced disability, and lumbar fatigability. Spine (Phila Pa 1976) 1999;24:1034–42.

Keijsers J, Steenbakkers M, Meertens R, et al. The Efficacy of the Back School: A Randomized Trial. Arthritis Care and Research 1990 December;3(4):204–9.

Keijsers JF, Groenman NH, Gerards FM, et al. A back school in The Netherlands: evaluating the results. Patient Educ Couns 1989;14:31–44.

Kellett KM, Kellett DA, NordholmLA. Effects of an exercise program on sick leave due to back pain. Phys Ther 1991;71:283–91; discussion 291-3.

Kerr DP, Walsh DM, Baxter D. Acupuncture in the management of chronic low back pain: a blinded randomized controlled trial. Clin J Pain 2003;19:364–70.

Kinalski R, Kuwik W, Pietrzak D. The comparison of the results of manual therapy versus physiotherapy methods used in treatment of patients with low back pain syndromes. J Manual Medicine 1989;4:44–6.

Kittang G, Melvaer T, Baerheim A. [Acupuncture contra antiphlogistics in acute lumbago]. Tidsskr Nor Laegeforen 2001;121:1207–10.

Klaber Moffett JA, Chase SM, Portek I, et al. A controlled, prospective study to evaluate the effectiveness of a back school in the relief of chronic low back pain. Spine (Phila Pa 1976) 1986;11:120–2.

Klinger NM, Wilson RR, Kanniainen CM, et al. Intravenous orphenadrine for the treatment of lumbar paravertebral muscle pain. Current Therapeutic Research 1988;43:247–54.

Koes BW, Bouter LM, van Mameren H, et al. Randomised clinical trial of manipulative therapy and physiotherapy for persistent back and neck complaints: results of one year follow up. BMJ 1992;304:601–5.

Kole-Snijders AM, Vlaeyen JW, Goossens ME, et al. Chronic low-back pain: what does cognitive coping skills training add to operant behavioral treatment? Results of a randomized clinical trial. J Consult Clin Psychol 1999;67:931–44.

Kovacs FM, Abraira V, Pozo F, et al. Local and remote sustained trigger point therapy for exacerbations of chronic low back pain. A randomized, double-blind, controlled, multicenter trial. Spine (Phila Pa 1976) 1997;22:786–97.

Kovacs FM, Llobera J, Abraira V, et al. Effectiveness and cost-effectiveness analysis of neuroreflexotherapy for subacute and chronic low back pain in routine general practice: a cluster randomized, controlled trial. Spine (Phila Pa 1976) 2002;27:1149–59.

Lacey PH, Dodd GD, Shannon DJ. A double blind, placebo controlled study of piroxicam in the management of acute musculoskeletal disorders. Eur J Rheumatol Inflamm 1984;7:95–104.

Lankhorst GJ, Van de Stadt RJ, Vogelaar TW, et al. The effect of the Swedish Back School in chronic idiopathic low back pain. A prospective controlled study. Scand J Rehabil Med 1983;15:141–5.

Leclaire R, Esdaile JM, Suissa S, et al. Back school in a first episode of compensated acute low back pain: a clinical trial to assess efficacy and prevent relapse. Arch Phys Med Rehabil 1996;77:673–9.

Lehmann TR, Russell DW, Spratt KF. The impact of patients with nonorganic physical findings on a controlled trial of transcutaneous electrical nerve stimulation and electroacupuncture. Spine (Phila Pa 1976) 1983;8:625–34.

Leibing E, Leonhardt U, Koster G, et al. Acupuncture treatment of chronic low-back pain -- a randomized, blinded, placebo-controlled trial with 9-month follow-up. Pain 2002;96:189–96.

Lepisto P. A comparative trial of DS 103-282 and placebo in the treatment of acute skeletal spasms due to disorders of the back. Current Therapeutic Research 1979;26:454–9.

Lidstrom A, Zachrisson M. Physical therapy on low back pain and sciatica. An attempt at evaluation. Scand J Rehabil Med 1970;2:37–42.

Lie H, Frey S. [Mobilizing or stabilizing exercise in degenerative disk disease in the lumbar region?]. Tidsskr Nor Laegeforen 1999;119:2051–3.

Lindequist S, Lundberg B, Wikmark R, et al. Information and regime at low back pain. Scand J Rehabil Med 1984;16:113–6.

Lindstrom I, Ohlund C, Eek C, et al. The effect of graded activity on patients with subacute low back pain: a randomized prospective clinical study with an operant-conditioning behavioral approach. Phys Ther 1992;72:279–90; discussion 291-3.

Linton SJ, Bradley LA, Jensen I, et al. The secondary prevention of low back pain: a controlled study with follow-up. Pain 1989;36:197–207.

Listrat V, Dougados M, Chevalier X, et al. Comparison of the analgesic effect of Tenoxicam after oral or intramuscular administration. Drug Invest Suppl 1990;2:51–2.

Ljunggren AE, Weber H, Kogstad O, et al. Effect of exercise on sick leave due to low back pain. A randomized, comparative, long-term study. Spine (Phila Pa 1976) 1997;22:1610–6; discussion 1617.

Lonn JH, Glomsrod B, Soukup MG, et al. Active back school: prophylactic management for low back pain. A randomized, controlled, 1-year follow-up study. Spine (Phila Pa 1976) 1999;24:865–71.

MacDonald RS, Bell CM. An open controlled assessment of osteopathic manipulation in nonspecific low-back pain. Spine (Phila Pa 1976) 1990;15:364–70.

Malmivaara A, Hakkinen U, Aro T, et al. The treatment of acute low back pain--bed rest, exercises, or ordinary activity? N Engl J Med 1995;332:351–5.

Manniche C, Hesselsoe G, Bentzen L, et al. Clinical trial of intensive muscle training for chronic low back pain. Lancet 1988;2:1473–6.

Mannion AF, Muntener M, Taimela S, et al. A randomized clinical trial of three active therapies for chronic low back pain. Spine (Phila Pa 1976) 1999;24:2435–48.

Marchand S, Charest J, Li J, et al. Is TENS purely a placebo effect? A controlled study on chronic low back pain. Pain 1993;54:99–106.

Mathews JA, Mills SB, Jenkins VM, et al. Back pain and sciatica: controlled trials of manipulation, traction, sclerosant and epidural injections. Br J Rheumatol 1987;26:416–23.

Matsumo S, Kaneda K, Norhara Y. Clinical evaluation of ketoprofen (Orudis) in lumbago - a double-blind comparison with diclofenac sodium. Br J Clin Pract 1981;35:266.

Melzack R, Vetere P, Finch L. Transcutaneous electrical nerve stimulation for low back pain. A comparison of TENS and massage for pain and range of motion. Phys Ther 1983;63:489–93.

Mencke VM, Wieden TE, Hoppe M, et al. Acupuncture of shoulder pain and low back pain. Two Prospective double-blind studies 1988;4: 204–215.

Mendelson G, Selwood TS, Kranz H, et al. Acupuncture treatment of chronic back pain. A double-blind placebo-controlled trial. Am J Med 1983;74:49-55.

Meng CF, Wang D, Ngeow J, et al. Acupuncture for chronic low back pain in older patients: a randomized, controlled trial. Rheumatology (Oxford) 2003;42:1508–17.

Milgrom C, Finestone A, Lev B, et al. Overexertional lumbar and thoracic back pain among recruits: a prospective study of risk factors and treatment regimens. J Spinal Disord 1993;6:187–93.

Million R, Nilsen KH, Jayson MI, et al. Evaluation of low back pain and assessment of lumbar corsets with and without back supports. Ann Rheum Dis 1981;40:449–54.

Moffett JK, Torgerson D, Bell-Syer S, et al. Randomised controlled trial of exercise for low back pain: clinical outcomes, costs, and preferences. BMJ 1999;319:279–83.

Moll W. [Therapy of acute lumbovertebral syndromes through optimal muscle relaxation using diazepam. Results of a double-blind study on 68 cases]. Med Welt 1973;24:1747–51.

Molsberger AF, Mau J, Pawelec DB, et al. Does acupuncture improve the orthopedic management of chronic low back pain--a randomized, blinded, controlled trial with 3 months follow up. Pain 2002;99:579–87.

Moore SR and Shurman J. Combined neuromuscular electrical stimulation and transcutaneous electrical nerve stimulation for treatment of chronic back pain: a double-blind, repeated measures comparison. Arch Phys Med Rehabil 1997;78:55–60.

Moseley L. Combined physiotherapy and education is efficacious for chronic low back pain. Aust J Physiother 2002;48:297–302.

Muckle DS. Flurbiprofen for the treatment of soft tissue trauma. Am J Med 1986;80:76–80.

Newton-John TR, Spence SH,Schotte D. Cognitive-behavioural therapy versus EMG biofeedback in the treatment of chronic low back pain. Behav Res Ther 1995;33:691–7.

Nicholas MK, Wilson PH, Goyen J. Comparison of cognitive-behavioral group treatment and an alternative non-psychological treatment for chronic low back pain. Pain 1992;48:339–47.

Nicholas MK, Wilson PH, Goyen J. Operant-behavioural and cognitive-behavioural treatment for chronic low back pain. Behav Res Ther 1991;29:225–38.

Niemisto L, Lahtinen-Suopanki T, Rissanen P, et al. A randomized trial of combined manipulation, stabilizing exercises, and physician consultation compared to physician consultation alone for chronic low back pain. Spine (Phila Pa 1976) 2003;28:2185–91.

Nouwen A. EMG biofeedback used to reduce standing levels of paraspinal muscle tension in chronic low back pain. Pain 1983;17:353–60.

Ongley MJ, Klein RG, Dorman TA, et al. A new approach to the treatment of chronic low back pain. Lancet 1987;2:143–6.

Orava S. Medical treatment of acute low back pain. Diflunisal compared with indomethacin in acute lumbago. Int J Clin Pharmacol Res 1986;6:45–51.

Pena M. Etodolac: analgesic effects in musculoskeletal and postoperative pain. Rheumatol Int 1990;10 Suppl:9–16.

Penttinen J, Nevala-Puranen N, Airaksinen O, et al. Randomized controlled trial of back school with and without peer support. J Occup Rehabil 2002;12:21–9.

Petersen T, Kryger P, Ekdahl C, et al. The effect of McKenzie therapy as compared with that of intensive strengthening training for the treatment of patients with subacute or chronic low back pain: A randomized controlled trial. Spine (Phila Pa 1976) 2002;27:1702–9.

Pipino F, Menarini C, Lombardi G, et al. A direct Myotonolytic (Pridinol Mesilate) for the managemnt of chronic low back pain: A multicentre, comparitive clinical evaluation. European Journal of Clinical Research 1991;1:55–70.

Pope MH, Phillips RB, Haugh LD, et al. A prospective randomized three-week trial of spinal manipulation, transcutaneous muscle stimulation, massage and corset in the treatment of subacute low back pain. Spine (Phila Pa 1976) 1994;19:2571–7.

Postacchini F, Facchini M, and Palieri P. Efficacy of Various Forms of Conservative Treatment in Low Back Pain: A Comparitive Study. Neuro-Orthopedics 1988;6:28–35.

Pratzel HG, Alken RG, Ramm S. Efficacy and tolerance of repeated oral doses of tolperisone hydrochloride in the treatment of painful reflex muscle spasm: results of a prospective placebo-controlled double-blind trial. Pain 1996;67:417–25.

Preyde M. Effectiveness of massage therapy for subacute low-back pain: a randomized controlled trial. CMAJ 2000;162:1815–20.

Rasmussen-Barr E, Nilsson-Wikmar L, Arvidsson I. Stabilizing training compared with manual treatment in sub-acute and chronic low-back pain. Man Ther 2003;8:233–41.

Rittweger J, Just K, Kautzsch K, et al. Treatment of chronic lower back pain with lumbar extension and whole-body vibration exercise: a randomized controlled trial. Spine (Phila Pa 1976) 2002;27:1829–34.

Rose MJ, Reilly JP, Pennie B, et al. Chronic low back pain rehabilitation programs: a study of the optimum duration of treatment and a comparison of group and individual therapy. Spine (Phila Pa 1976) 1997;22:2246–51; discussion 2252–3.

Salzmann E, Pforringer W, Paal G, et al. Treatment of chronic low-back syndrome with tetrazepam in a placebo controlled double-blind trial. J. Drug Dev 1992;4:219–28.

Seferlis T, Nemeth G, Carlsson AM, et al. Conservative treatment in patients sick-listed for acute low-back pain: a prospective randomised study with 12 months' follow-up. Eur Spine J 1998;7:461–70.

Siegmeth W, Sieberer W. A comparison of the short-term effects of ibuprofen and diclofenac in spondylosis. J Int Med Res 1978;6:369–74.

Sims-Williams H, Jayson MI, Young SM, et al. Controlled trial of mobilisation and manipulation for low back pain: hospital patients. Br Med J 1979;2:1318–20.

Soukup MG, Glomsrod B, Lonn J H, et al. The effect of a Mensendieck exercise program as secondary prophylaxis for recurrent low back pain. A randomized, controlled trial with 12-month follow-up. Spine (Phila Pa 1976), 1999;24(15):1585–91.

Srini NV, Duncan T. Femoral hernia in children. Br J Clin Pract 1987;41:618.

Staal JB, Hlobil H, Twisk JW, et al. Graded activity for low back pain in occupational health care: a randomized, controlled trial. Ann Intern Med 2004;140:77–84.

Stankovic R, Johnell O. Conservative treatment of acute low-back pain. A prospective randomized trial: McKenzie method of treatment versus patient education in "mini back school." Spine (Phila Pa 1976) 1990;15:120–3.

Storheim K, Brox JI, Holm I, et al. Intensive group training versus cognitive intervention in sub-acute low back pain: short-term results of a single-blind randomized controlled trial. J Rehabil Med 2003;35:132–40.

Stratz T. [Intramuscular etofenamate in the treatment of acute lumbago. Effectiveness and tolerance in comparison with intramuscular diclofenac-Na]. Fortschr Med 1990;108:264–6.

Stuckey SJ, Jacobs A, Goldfarb J. EMG biofeedback training, relaxation training, and placebo for the relief of chronic back pain. Percept Mot Skills 1986;63:1023–36.

Sweetman BJ, Baig A, Parsons DL. Mefenamic acid, chlormezanone-paracetamol, ethoheptazine-aspirin-meprobamate: a comparative study in acute low back pain. Br J Clin Pract. 1987 Feb;41(2):619–24.

Szpalski M, Hayez JP. How many days of bed rest for acute low back pain? Objective assessment of trunk function. Eur Spine J 1992;1:29–31.

Szpalski M, Hayez JP. Objective functional assessment of the efficacy of tenoxicam in the treatment of acute low back pain. A double-blind placebo-controlled study. Br J Rheumatol 1994;33:74–8.

Szpalski M, Poty S, Hayez JP, et al. Objective assessment of trunk function in patients with acute low back pain treated with Tenoxicam A prospective controlled study. Neuro-Ortopedics 1990;10:41–47.

Tervo T, Petaja L, Lepisto P. A controlled clinical trial of a muscle relaxant analgesic combination in the treatment of acute lumbago. Br J Clin Pract 1976;30:62–4.

Thomas M, Lundberg T. Importance of modes of acupuncture in the treatment of chronic nociceptive low back pain. Acta Anaesthesiol Scand 1994;38:63–9.

Timm KE. A randomized-control study of active and passive treatments for chronic low back pain following L5 laminectomy. J Orthop Sports Phys Ther 1994;20:276–86.

Torstensen TA, Ljunggren AE, Meen HD, et al. Efficiency and costs of medical exercise therapy, conventional physiotherapy, and self-exercise in patients with chronic low back pain. A pragmatic, randomized, single-blinded, controlled trial with 1-year follow-up. Spine (Phila Pa 1976) 1998;23:2616–24.

Triano JJ, McGregor M, Hondras MA, et al. Manipulative therapy versus education programs in chronic low back pain. Spine (Phila Pa 1976) 1995;20:948–55.

Tritilanunt T, Wajanavisit W. The efficacy of an aerobic exercise and health education program for treatment of chronic low back pain. J Med Assoc Thai 2001;84 Suppl 2:S528–33.

Tsukayama H, Yamashita H, Amagai H, et al. Randomised controlled trial comparing the effectiveness of electroacupuncture and TENS for low back pain: a preliminary study for a pragmatic trial. Acupunct Med 2002;20:175–80.

Turner JA. Comparison of group progressive-relaxation training and cognitive-behavioral group therapy for chronic low back pain. J Consult Clin Psychol 1982;50:757–65.

Turner JA, Clancy S. Comparison of operant behavioral and cognitive-behavioral group treatment for chronic low back pain. J Consult Clin Psychol 1988;56:261–6.

Turner JA, Clancy S, McQuade KJ, et al. Effectiveness of behavioral therapy for chronic low back pain: a component analysis. J Consult Clin Psychol 1990;58:573–9.

Turner JA, Jensen MP. Efficacy of cognitive therapy for chronic low back pain. Pain 1993;52:169–77.

Underwood MR, Morgan J. The use of a back class teaching extension exercises in the treatment of acute low back pain in primary care. Fam Pract 1998;15:9–15.

Valle-Jones JC, Walsh H, O'Hara J, et al. Controlled trial of a back support ('Lumbotrain') in patients with non-specific low back pain. Curr Med Res Opin 1992;12:604–13.

Van den Hout JH, Vlaeyen JW, Heuts PH, et al. Secondary prevention of work-related disability in nonspecific low back pain: does problem-solving therapy help? A randomized clinical trial. Clin J Pain 2003;19:87–96.

Vetter G, Bruggemann G, Lettko M, et al. [Shortening diclofenac therapy by B vitamins. Results of a randomized double-blind study, diclofenac 50 mg versus diclofenac 50 mg plus B vitamins, in painful spinal diseases with degenerative changes]. Z Rheumatol 1988;47:351–62.

Videman T, Heikkila J, Partanen T. Double-blind parallel study of meptazinol versus diflunisal in the treatment of lumbago. Curr Med Res Opin 1984;9:246–52.

Videman T, Osterman K. Double-blind parallel study of piroxicam versus indomethacin in the treatment of low back pain. Ann Clin Res 1984;16:156–60.

Vroomen PC, de Krom MC, Wilmink JT, et al. Lack of effectiveness of bed rest for sciatica. N Engl J Med 1999;340:418–23.

Waagen GN, Haldeman S, Cook G, et al. Short term trial of chiropractic adjustments for the relief of chronic low back pain. Manual Medicine 1986;2:63–7.

Waikakul S, Danputipong P, Soparat K. Topical analgesics, indomethacin plaster and diclofenac emulgel for low back pain: a parallel study. J Med Assoc Thai 1996;79;486–90.

Waikakul S, Soparat K. Effectiveness and Safety of Loxoprofen Compared with Naproxen in Nonsurgical Low Back Pain. A Parallel Study. Clin Drug Invest 1995;10:59–63.

Wang RR, Tronnier V. Effect of acupuncture on pain management in patients before and after lumbar disc protrusion surgery--a randomized control study. Am J Chin Med 2000;28:25–33.

Waterworth RF, Hunter IA. An open study of diflunisal, conservative and manipulative therapy in the management of acute mechanical low back pain. N Z Med J 1985;98:372–5.

Weber H. Comparison of the effect of diazepam and levomepromazine on pain in patients with acute lumbago-sciatica. J Oslo City Hosp 1980;30:65–8.

Weber H, Aasand G. The effect of phenylbutazone on patients with acute lumbago-sciatica. A double blind trial. J Oslo City Hosp 1980;30:69–72.

Weber H, Holme I, Amlie E. The natural course of acute sciatica with nerve root symptoms in a double-blind placebo-controlled trial evaluating the effect of piroxicam. Spine (Phila Pa 1976) 1993;18:1433–8.

Wilkinson MJ. Does 48 hours' bed rest influence the outcome of acute low back pain? Br J Gen Pract 1995;45:481–4.

Worz R, Bolten W, Heller B, et al. [Flupirtine in comparison with chlormezanone in chronic musculoskeletal back pain. Results of a multicenter randomized double-blind study]. Fortschr Med 1996;114:500–4.

Wreje U, Nordgren B, Aberg H. Treatment of pelvic joint dysfunction in primary care--a controlled study. Scand J Prim Health Care 1992;10:310–5.

Yamashita H. Are the effects of electro-acupuncture on low back pain equal to those of TENS? Focus on Alternative and Complementary Therapies 2001;6:254–5.

Yelland MJ, Glasziou PP, Bogduk N, et al. Prolotherapy injections, saline injections, and exercises for chronic low-back pain: a randomized trial. Spine (Phila Pa 1976) 2004;29:9–16; discussion 16.

Yeung CK, Leung MC, Chow DH. The use of electro-acupuncture in conjunction with exercise for the treatment of chronic low-back pain. J Altern Complement Med 2003;9:479–90.

Yozbatiran N, Yildirim Y, Parlak B. Effects of fitness and aquafitness exercises on physical fitness in patients with chronic low back pain. The Pain Clinic 2004;16:35–42.

Appendix B. References Dataset 2: EPC Reports, 165 Trials

[no author] Gabapentin in partial epilepsy. UK Gabapentin Study Group. Lancet 1990 May 12;335(8698):1114–7.

[no author] Efficacy of felbamate in childhood epileptic encephalopathy (Lennox-Gastaut syndrome). The Felbamate Study Group in Lennox-Gastaut Syndrome. N Engl J Med 1993 Jan 7;328(1):29–33.

[no author] Dietary supplementation with n-3 polyunsaturated fatty acids and vitamin E after myocardial infarction: results of the GISSI-Prevenzione trial. Gruppo Italiano per lo Studio della Sopravvivenza nell'Infarto miocardico. Lancet 1999 Aug 7;354(9177):447–55.

[no author] Topiramate in medically intractable partial epilepsies: double-blind placebo-controlled randomized parallel group trial. Korean Topiramate Study Group. Epilepsia 1999 Dec;40(12):1767–74.

Altman RD, Moskowitz R. Intraarticular sodium hyaluronate (Hyalgan) in the treatment of patients with osteoarthritis of the knee: a randomized clinical trial. Hyalgan Study Group. J Rheumatol 1998 Nov;25(11):2203–12.

Ancarani E, Biondi B, Bolletta A, Cestra D, De Bella E, Nirchi M, et al. Major depression complicating hemodialysis in patients with chronic renal failure: A multicenter, double-blind, controlled clinical trial of s-adenosyl-l-methionine versus placebo. Current Therapeutic Research 1993;54(6):680–6.

Anderson RA, Cheng N, Bryden NA, Polansky MM, Chi J, Feng J. Elevated intakes of supplemental chromium improve glucose and insulin variables in individuals with type 2 diabetes. Diabetes 1997 Nov;46(11):1786–91.

Angerer P, Kothny W, Stork S, von Schacky C. Effect of dietary supplementation with omega-3 fatty acids on progression of atherosclerosis in carotid arteries. Cardiovasc Res 2002 Apr;54(1):183–90.

Anhut H, Ashman P, Feuerstein TJ, Sauermann W, Saunders M, Schmidt B. Gabapentin (Neurontin) as add-on therapy in patients with partial seizures: a double-blind, placebo-controlled study. The International Gabapentin Study Group. Epilepsia 1994 Jul-Aug;35(4):795–801.

Antoni CE, Kavanaugh A, Kirkham B, Tutuncu Z, Burmester GR, Schneider U, et al. Sustained benefits of infliximab therapy for dermatologic and articular manifestations of psoriatic arthritis: results from the infliximab multinational psoriatic arthritis controlled trial (IMPACT). Arthritis Rheum 2005 Apr;52(4):1227–36.

Appleton R, Fichtner K, LaMoreaux L, Alexander J, Halsall G, Murray G, et al. Gabapentin as add-on therapy in children with refractory partial seizures: a 12-week, multicentre, double-blind, placebo-controlled study. Gabapentin Paediatric Study Group. Epilepsia 1999 Aug;40(8):1147–54.

Atmaca M, Kuloglu M, Tezcan E, Gecici O. Quetiapine augmentation in patients with treatment resistant obsessive-compulsive disorder: a single-blind, placebo-controlled study. Int Clin Psychopharmacol 2002 May;17(3):115–9.

Bairati I, Roy L, Meyer F. Effects of a fish oil supplement on blood pressure and serum lipids in patients treated for coronary artery disease. Can J Cardiol 1992 Jan-Feb;8(1):41–6.

Bakris G, Calhoun D, Egan B, Hellmann C, Dolker M, Kingma I. Orlistat improves blood pressure control in obese subjects with treated but inadequately controlled hypertension. J Hypertens 2002 Nov;20(11):2257–67.

Ben-Menachem E, Falter U. Efficacy and tolerability of levetiracetam 3000 mg/d in patients with refractory partial seizures: a multicenter, double-blind, responder-selected study evaluating monotherapy. European Levetiracetam Study Group. Epilepsia 2000 Oct;41(10):1276–83.

Ben-Menachem E, Henriksen O, Dam M, Mikkelsen M, Schmidt D, Reid S, et al. Double-blind, placebo-controlled trial of topiramate as add-on therapy in patients with refractory partial seizures. Epilepsia 1996 Jun;37(6):539–43.

Betts T, Waegemans T, Crawford P. A multicentre, double-blind, randomized, parallel group study to evaluate the tolerability and efficacy of two oral doses of levetiracetam, 2000 mg daily and 4000 mg daily, without titration in patients with refractory epilepsy. Seizure 2000 Mar;9(2):80–7.

Biton V, Montouris GD, Ritter F, Riviello JJ, Reife R, Lim P, et al. A randomized, placebo-controlled study of topiramate in primary generalized tonic-clonic seizures. Topiramate YTC Study Group. Neurology 1999 Apr 22;52(7):1330–7.

Bonaa KH, Bjerve KS, Nordoy A. Docosahexaenoic and eicosapentaenoic acids in plasma phospholipids are divergently associated with high density lipoprotein in humans. Arterioscler Thromb 1992 Jun;12(6):675–81.

Brandt KD, Block JA, Michalski JP, Moreland LW, Caldwell JR, Lavin PT. Efficacy and safety of intraarticular sodium hyaluronate in knee osteoarthritis. ORTHOVISC Study Group. Clin Orthop Relat Res 2001 Apr(385):130–43.

Brodaty H, Corey-Bloom J, Potocnik FC, Truyen L, Gold M, Damaraju CR. Galantamine prolonged-release formulation in the treatment of mild to moderate Alzheimer's disease. Dement Geriatr Cogn Disord 2005;20(2-3):120–32.

Broom I, Wilding J, Stott P, Myers N. Randomised trial of the effect of orlistat on body weight and cardiovascular disease risk profile in obese patients: UK Multimorbidity Study. Int J Clin Pract 2002 2002 Sep;56(7):494–9.

Brown KM, Morrice PC, Duthie GG. Vitamin E supplementation suppresses indexes of lipid peroxidation and platelet counts in blood of smokers and nonsmokers but plasma lipoprotein concentrations remain unchanged. Am J Clin Nutr 1994 Sep;60(3):383–7.

Brown SA, Garcia AA, Kouzekanani K, Hanis CL. Culturally competent diabetes self-management education for Mexican Americans: the Starr County border health initiative. Diabetes Care 2002 Feb;25(2):259–68.

Brox J, Olaussen K, Osterud B, Elvevoll EO, Bjornstad E, Brattebog G, et al. A long-term seal- and cod-liver-oil supplementation in hypercholesterolemic subjects. Lipids 2001 Jan;36(1):7–13.

Bystritsky A, Ackerman DL, Rosen RM, Vapnik T, Gorbis E, Maidment KM, et al. Augmentation of serotonin reuptake inhibitors in refractory obsessive-compulsive disorder using adjunctive olanzapine: a placebo-controlled trial. J Clin Psychiatry 2004 Apr;65(4):565–8.

Cairns JA, Gill J, Morton B, Roberts R, Gent M, Hirsh J, et al. Fish oils and low-molecular-weight heparin for the reduction of restenosis after percutaneous transluminal coronary angioplasty. The EMPAR Study. Circulation 1996 Oct 1;94(7):1553–60.

Carey PD, Vythilingum B, Seedat S, Muller JE, van Ameringen M, Stein DJ. Quetiapine augmentation of SRIs in treatment refractory obsessive-compulsive disorder: a double-blind, randomised, placebo-controlled study [ISRCTN83050762]. BMC Psychiatry 2005;5:5.

Carney MW, Edeh J, Bottiglieri T, Reynolds EM, Toone BK. Affective illness and S-adenosyl methionine: a preliminary report. Clin Neuropharmacol 1986;9(4):379–85.

Carrabba M, Paresce E, Angelini M, Re KA, Tochiana EEM, Perbellini A. The safety and efficacy of different dose schedules of hyaluronic acid in the treatment of painful osteoarthritis of the knee with joint effusion Eur J Rheumatol Inflamm 1995;15:25–31.

Carrieri P, Indaco A, Gentile S, Troisi E, Campanella G. S-adenosylmethionine treatment of depression in patients with parkinson's disease. Current Therapeutic Research 1990;48(1):154–60.

Caruso I, Fumagaili M, Boccassini L, Sarzi Puttin P, Santandrea S, Ciniseili G, et al. Treatment of depression in rheumatoid arthritic patients: A comparison of S-Adenosylmethinine (Samyr*) and placebo in a double-blind study. Clinical Trials Journal 1987;24(4):305–10.

Cereghino JJ, Biton V, Abou-Khalil B, Dreifuss F, Gauer LJ, Leppik I. Levetiracetam for partial seizures: results of a double-blind, randomized clinical trial. Neurology 2000 Jul 25;55(2):236–42.

Chadwick D, Leiderman DB, Sauermann W, Alexander J, Garofalo E. Gabapentin in generalized seizures. Epilepsy Res 1996 Nov;25(3):191–7.

Cohen MA, Shiroky JB, Ballechey ML, Neville C, Esdaile JM. Double-blind randomized trial of Intra-articular Hyaluronate in the treatment of osteoarthritis of the knee. Arthritis Rheum 1994;34(6).

Corey-Bloom J, Anand JV, Veach J. A randomized trial evaluating the efficacy and safety of ENA 713 (rivastigmine tartrate), a new acetylcholinesterase inhibitor, in patients with mild to moderately severe Alzheimer's disease. Int J Geriatr Psychopharmacol 1998;1:55–65.

Corrado E, Peluso GF, Gigliotti S. The effects of intra-articular administrationof hyaluronic acid on osteoarthritis of the knee: a clinical study with immunological and biochemical evaluations Eur J Rheumatol Inflamm 1995;15:47–56.

Creamer P, Sharif M, George E, Meadows K, Cushnaghan J, Shinmei M, et al. Intra-articular hyaluronic acid in osteoarthritis of the knee: an investigation into mechanisms of action. Osteoarthritis Cartilage 1994 Jun;2(2):133–40.

Dahlberg L, Lohmander LS, Ryd L. Intraarticular injections of hyaluronan in patients with cartilage abnormalities and knee pain. A one-year double-blind, placebo-controlled study. Arthritis Rheum 1994 Apr;37(4):521–8.

Davidson M, Hauptman J, DiGirolamo M, Foreyt J, Halsted C, Heber D, et al. Weight control and risk factor reduction in obese subjects treated for 2 years with orlistat: a randomized controlled trial. JAMA 1999 Jan 20;281(3):235–42.

Davidson MB, Castellanos M, Kain D, Duran P. The effect of self monitoring of blood glucose concentrations on glycated hemoglobin levels in diabetic patients not taking insulin: a blinded, randomized trial. Am J Med 2005 Apr;118(4):422–5.

De Leo D. S-adenosylmethionine as an antidepressant: A double-blind trial versus placebo. Current Therapeutic Research 1987;41(6):865–70.

de Waart FG, Moser U, Kok FJ. Vitamin E supplementation in elderly lowers the oxidation rate of linoleic acid in LDL. Atherosclerosis 1997 Sep;133(2):255–63.

Delle Chiaie R, Boissard G. Ademetionine (same) for the treatment of major depression: Meta-analysis of two european multicenter controlled trials. Pharmacol Res 1997;35(108):108.

DeMaio SJ, King SB, 3rd, Lembo NJ, Roubin GS, Hearn JA, Bhagavan HN, et al. Vitamin E supplementation, plasma lipids and incidence of restenosis after percutaneous transluminal coronary angioplasty (PTCA). J Am Coll Nutr 1992 Feb;11(1):68–73.

Denys D, de Geus F, van Megen HJ, Westenberg HG. A double-blind, randomized, placebo-controlled trial of quetiapine addition in patients with obsessive-compulsive disorder refractory to serotonin reuptake inhibitors. J Clin Psychiatry 2004 Aug;65(8):1040–8.

D'Eramo-Melkus GA, Wylie-Rosett J, Hagan JA. Metabolic impact of education in NIDDM. Diabetes Care 1992 Jul;15(7):864–9.

Derosa G, Mugellini A, Ciccarelli L, Fogari R. Randomized, double-blind, placebo-controlled comparison of the action of orlistat, fluvastatin, or both an anthropometric measurements, blood pressure, and lipid profile in obese patients with hypercholesterolemia prescribed a standardized diet. Clin Ther 2003 2003 Apr;25(4):1107–22.

Dixon AS, Jacoby RK, Berry H, Hamilton EB. Clinical trial of intra-articular injection of sodium hyaluronate in patients with osteoarthritis of the knee. Curr Med Res Opin 1988;11(4):205–13.

Dougados M, Nguyen M, Listrat V, Amor B. High molecular weight sodium hyaluronate (hyalectin) in osteoarthritis of the knee: a 1 year placebo-controlled trial. Osteoarthritis Cartilage 1993 Apr;1(2):97–103.

Duchowny M, Pellock JM, Graf WD, Billard C, Gilman J, Casale E, et al. A placebo-controlled trial of lamotrigine add-on therapy for partial seizures in children. Lamictal Pediatric Partial Seizure Study Group. Neurology 1999 Nov 10;53(8):1724–31.

Elterman RD, Glauser TA, Wyllie E, Reife R, Wu SC, Pledger G. A double-blind, randomized trial of topiramate as adjunctive therapy for partial-onset seizures in children. Topiramate YP Study Group. Neurology 1999 Apr 22;52(7):1338–44.

Eritsland J, Arnesen H, Seljeflot I, Hostmark AT. Long-term metabolic effects of n-3 polyunsaturated fatty acids in patients with coronary artery disease. Am J Clin Nutr 1995 Apr;61(4):831–6.

Erzegovesi S, Guglielmo E, Siliprandi F, Bellodi L. Low-dose risperidone augmentation of fluvoxamine treatment in obsessive-compulsive disorder: a double-blind, placebo-controlled study. Eur Neuropsychopharmacol 2005 Jan;15(1):69–74.

Evans GW. The effect of chromium picolinate on insulin controlled parameters in humans. Int J Biosoc Med Res 1989;11:163–80.

Falkenberg MG, Elwing BE, Goransson AM, Hellstrand BE, Riis UM. Problem oriented participatory education in the guidance of adults with non-insulin-treated type-II diabetes mellitus. Scand J Prim Health Care 1986 Sep;4(3):157–64.

Farmer A, Wade A, Goyder E, Yudkin P, French D, Craven A, et al. Impact of self monitoring of blood glucose in the management of patients with non-insulin treated diabetes: open parallel group randomised trial. BMJ 2007 Jul 21;335(7611):132.

Faught E, Ayala R, Montouris GG, Leppik IE. Randomized controlled trial of zonisamide for the treatment of refractory partial-onset seizures. Neurology 2001 Nov 27;57(10):1774–9.

Faught E, Wilder BJ, Ramsay RE, Reife RA, Kramer LD, Pledger GW, et al. Topiramate placebo-controlled dose-ranging trial in refractory partial epilepsy using 200-, 400-, and 600-mg daily dosages. Topiramate YD Study Group. Neurology 1996 Jun;46(6):1684–90.

Fava M, Rosenbaum JF, Birnbaum R, Kelly K, Otto MW, MacLaughlin R. The thyrotropin response to thyrotropin-releasing hormone as a predictor of response to treatment in depressed outpatients. Acta Psychiatr Scand 1992 Jul;86(1):42–5.

Fineberg NA, Sivakumaran T, Roberts A, Gale T. Adding quetiapine to SRI in treatment-resistant obsessive-compulsive disorder: a randomized controlled treatment study. Int Clin Psychopharmacol 2005 Jul;20(4):223–6.

Finnegan YE, Minihane AM, Leigh-Firbank EC, Kew S, Meijer GW, Muggli R, et al. Plant- and marine-derived n-3 polyunsaturated fatty acids have differential effects on fasting and postprandial blood lipid concentrations and on the susceptibility of LDL to oxidative modification in moderately hyperlipidemic subjects. Am J Clin Nutr 2003 Apr;77(4):783–95.

Fontbonne A, Billault B, Acosta M, Percheron C, Varenne P, Besse A, et al. Is glucose self-monitoring beneficial in non-insulin-treated diabetic patients? Results of a randomized comparative trial. Diabete Metab 1989 Sep–Oct;15(5):255–60.

Franzen D, Schannwell M, Oette K, Hopp HW. A prospective, randomized, and double-blind trial on the effect of fish oil on the incidence of restenosis following PTCA. Cathet Cardiovasc Diagn 1993 Apr;28(4):301–10.

Fuller CJ, Chandalia M, Garg A, Grundy SM, Jialal I. RRR-alpha-tocopheryl acetate supplementation at pharmacologic doses decreases low-density-lipoprotein oxidative susceptibility but not protein glycation in patients with diabetes mellitus. Am J Clin Nutr 1996 May;63(5):753–9.

Ghosh D, Bhattacharya B, Mukherjee B, Manna B, Sinha M, Chowdhury J, et al. Role of chromium supplementation in Indians with type 2 diabetes mellitus. J Nutr Biochem 2002 Nov;13(11):690–7.

Glasgow RE, Toobert DJ, Hampson SE, Brown JE, Lewinsohn PM, Donnelly J. Improving self-care among older patients with type II diabetes: the "Sixty Something..." Study. Patient Educ Couns 1992 Feb;19(1):61–74.

Glauser TA, Nigro M, Sachdeo R, Pasteris LA, Weinstein S, Abou-Khalil B, et al. Adjunctive therapy with oxcarbazepine in children with partial seizures. The Oxcarbazepine Pediatric Study Group. Neurology 2000 Jun 27;54(12):2237–44.

Goudswaard AN, Stolk RP, Zuithoff NP, de Valk HW, Rutten GE. Long-term effects of self-management education for patients with Type 2 diabetes taking maximal oral hypoglycaemic therapy: a randomized trial in primary care. Diabet Med 2004 May;21(5):491–6.

Greenfield S, Kaplan SH, Ware JE, Jr., Yano EM, Frank HJ. Patients' participation in medical care: effects on blood sugar control and quality of life in diabetes. J Gen Intern Med 1988 Sep-Oct;3(5):448–57.

Grimsgaard S, Bonaa KH, Hansen JB, Nordoy A. Highly purified eicosapentaenoic acid and docosahexaenoic acid in humans have similar triacylglycerol-lowering effects but divergent effects on serum fatty acids. Am J Clin Nutr 1997 Sep;66(3):649–59.

Guerci B, Drouin P, Grange V, Bougneres P, Fontaine P, Kerlan V, et al. Self-monitoring of blood glucose significantly improves metabolic control in patients with type 2 diabetes mellitus: the Auto-Surveillance Intervention Active (ASIA) study. Diabetes Metab 2003 Dec;29(6):587–94.

Hanefeld M, Sachse G. The effects of orlistat on body weight and glycaemic control in overweight patients with type 2 diabetes: a randomized, placebo-controlled trial. Diabetes Obes Metab 2002 Nov;4(6):415–23.

Hauptman J, Lucas C, Boldrin M, Collins H, Segal K. Orlistat in the long-term treatment of obesity in primary care settings. Arch Fam Med 2000 Feb;9(2):160–7.

Henderson EB, Smith EC, Pegley F, Blake DR. Intra-articular injections of 750 kD hyaluronan in the treatment of osteoarthritis: a randomised single centre double-blind placebo-controlled trial of 91 patients demonstrating lack of efficacy. Ann Rheum Dis 1994 Aug;53(8):529–34.

Hill J, Hauptman J, Anderson J, Fujioka K, O'Neil P, Smith D, et al. Orlistat, a lipase inhibitor, for weight maintenance after conventional dieting: a 1-y study. Am J Clin Nutr 1999 Jun;69(6):1108–16.

Hoffman RM, Garewal HS. Alpha-tocopherol supplementation for men with existing coronary artery disease: a feasibility study. Prev Med 1999 Aug;29(2):112–8.

Hollander E, Baldini Rossi N, Sood E, Pallanti S. Risperidone augmentation in treatment-resistant obsessive-compulsive disorder: a double-blind, placebo-controlled study. Int J Neuropsychopharmacol 2003 Dec;6(4):397–401.

Hollander P, Elbein S, Hirsch I, Kelley D, McGill J, Taylor T, et al. Role of orlistat in the treatment of obese patients with type 2 diabetes. A 1-year randomized double-blind study. Diabetes Care 1998 1998 Aug;21(8):1288–94.

Homma A, Takeda M, Imai Y, Udaka F, Hasegawa K, Kameyama M, et al. Clinical efficacy and safety of donepezil on cognitive and global function in patients with Alzheimer's disease. A 24-week, multicenter, double-blind, placebo-controlled study in Japan. E2020 Study Group. Dement Geriatr Cogn Disord 2000 Nov-Dec;11(6):299–313.

Huskisson EC, Donnelly S. Hyaluronic acid in the treatment of osteoarthritis of the knee. Rheumatology (Oxford) 1999 Jul;38(7):602–7.

Jaber LA, Halapy H, Fernet M, Tummalapalli S, Diwakaran H. Evaluation of a pharmaceutical care model on diabetes management. Ann Pharmacother 1996 Mar;30(3):238–43.

Jaber LA, Halapy H, Fernet M, Tummalapalli S, Diwakaran H. Evaluation of a pharmaceutical care model on diabetes management. Ann Pharmacother 1996 Mar;30(3):238–43.

Jain SK, McVie R, Jaramillo JJ, Palmer M, Smith T, Meachum ZD, et al. The effect of modest vitamin E supplementation on lipid peroxidation products and other cardiovascular risk factors in diabetic patients. Lipids 1996 Mar;31 Suppl:S87–90.

Jennings PE, Morgan HC, Barnett AH. Improved diabetes control and knowledge during a diabetes self-help group. Diabetes Educ 1987 Fall;13(4):390–3.

Jialal I, Fuller CJ, Huet BA. The effect of alpha-tocopherol supplementation on LDL oxidation. A dose-response study. Arterioscler Thromb Vasc Biol 1995 Feb;15(2):190–8.

Jialal I, Grundy SM. Effect of dietary supplementation with alpha-tocopherol on the oxidative modification of low density lipoprotein. J Lipid Res 1992 Jun;33(6):899–906.

Jubb RW, Piva S, Beinat L, Dacre J, Gishen P. A one-year, randomised, placebo (saline) controlled clinical trial of 500-730 kDa sodium hyaluronate (Hyalgan) on the radiological change in osteoarthritis of the knee. Int J Clin Pract 2003 Jul-Aug;57(6):467–74.

Kagan BL, Sultzer DL, Rosenlicht N, Gerner RH. Oral S-adenosylmethionine in depression: a randomized, double-blind, placebo-controlled trial. Am J Psychiatry 1990 May;147(5):591–5.

Kaikkonen J, Nyyssonen K, Tomasi A, Iannone A, Tuomainen TP, Porkkala-Sarataho E, et al. Antioxidative efficacy of parallel and combined supplementation with coenzyme Q10 and d-alpha-tocopherol in mildly hypercholesterolemic subjects: a randomized placebo-controlled clinical study. Free Radic Res 2000 Sep;33(3):329–40.

Karhunen L, Franssila-Kallunki A, Rissanen P, Valve R, Kolehmainen M, Rissanen A, et al. Effect of orlistat treatment on body composition and resting energy expenditure during a two-year weight-reduction programme in obese Finns. Int J Obes Relat Metab Disord 2000 Dec;24(12):1567–72.

Karlsson J, Sjogren LS, Lohmander LS. Comparison of two hyaluronan drugs and placebo in patients with knee osteoarthritis. A controlled, randomized, double-blind, parallel-design multicentre study. Rheumatology (Oxford) 2002 Nov;41(11):1240–8.

Kelley D, Bray G, Pi-Sunyer F, Klein S, Hill J, Miles J, et al. Clinical efficacy of orlistat therapy in overweight and obese patientswith insulin-treated type 2 diabetes: a 1-year randomized controlledtrial. Diabetes Care 2002;25(6):1033–41.

Keyserling TC, Samuel-Hodge CD, Ammerman AS, Ainsworth BE, Henriquez-Roldan CF, Elasy TA, et al. A randomized trial of an intervention to improve self-care behaviors of African-American women with type 2 diabetes: impact on physical activity. Diabetes Care 2002 Sep;25(9):1576–83.

Kibriya MG, Ali L, Banik NG, Khan AK. Home monitoring of blood glucose (HMBG) in Type-2 diabetes mellitus in a developing country. Diabetes Res Clin Pract 1999 Dec;46(3):253–7.

Kim HS, Oh JA. Adherence to diabetes control recommendations: impact of nurse telephone calls. J Adv Nurs 2003 Nov;44(3):256–61.

Kleefstra N, Houweling ST, Jansman FG, Groenier KH, Gans RO, Meyboom-de Jong B, et al. Chromium treatment has no effect in patients with poorly controlled, insulin-treated type 2 diabetes in an obese Western population: a randomized, double-blind, placebo-controlled trial. Diabetes Care 2006 Mar;29(3):521–5.

Krempf M, Louvet JP, Allanic H, Miloradovich T, Joubert JM, Attali JR. Weight reduction and long-term maintenance after 18 months treatment with orlistat for obesity. Int J Obes Relat Metab Disord 2003 May;27(5):591–7.

Kwon HS, Cho JH, Kim HS, Song BR, Ko SH, Lee JM, et al. Establishment of blood glucose monitoring system using the internet. Diabetes Care 2004 Feb;27(2):478–83.

Lee NA, Reasner CA. Beneficial effect of chromium supplementation on serum triglyceride levels in NIDDM. Diabetes Care 1994 Dec;17(12):1449–52.

Leigh-Firbank EC, Minihane AM, Leake DS, Wright JW, Murphy MC, Griffin BA, et al. Eicosapentaenoic acid and docosahexaenoic acid from fish oils: differential associations with lipid responses. Br J Nutr 2002 May;87(5):435–45.

Leng GC, Lee AJ, Fowkes FG, Jepson RG, Lowe GD, Skinner ER, et al. Randomized controlled trial of gamma-linolenic acid and eicosapentaenoic acid in peripheral arterial disease. Clin Nutr 1998 Dec;17(6):265–71.

Levetan CS, Dawn KR, Robbins DC, Ratner RE. Impact of computer-generated personalized goals on HbA(1c). Diabetes Care 2002 Jan;25(1):2–8.

Lindgarde F. The effect of orlistat on body weight and coronary heart disease risk profile in obese patients: the Swedish Multimorbidity Study. J Intern Med 2000 Sep;248(3):245–54.

Lohmander LS, Dalen N, Englund G, Hamalainen M, Jensen EM, Karlsson K, et al. Intra-articular hyaluronan injections in the treatment of osteoarthritis of the knee: a randomised, double blind, placebo controlled multicentre trial. Hyaluronan Multicentre Trial Group. Ann Rheum Dis 1996 Jul;55(7):424–31.

Lucas CP, Boldrin MN, Reaven GM. Effect of orlistat added to diet (30% of calories from fat) on plasma lipids, glucose, and insulin in obese patients with hypercholesterolemia. Am J Cardiol 2003 Apr 15;91(8):961–4.

Lungershausen YK, Abbey M, Nestel PJ, Howe PR. Reduction of blood pressure and plasma triglycerides by omega-3 fatty acids in treated hypertensives. J Hypertens 1994 Sep;12(9):1041–5.

Matsuo F, Bergen D, Faught E, Messenheimer JA, Dren AT, Rudd GD, et al. Placebo-controlled study of the efficacy and safety of lamotrigine in patients with partial seizures. U.S. Lamotrigine Protocol 0.5 Clinical Trial Group. Neurology 1993 Nov;43(11):2284–91.

McCulloch DK, Mitchell RD, Ambler J, Tattersall RB. Influence of imaginative teaching of diet on compliance and metabolic control in insulin dependent diabetes. Br Med J (Clin Res Ed) 1983 Dec 17;287(6408):1858–61.

McDougle CJ, Epperson CN, Pelton GH, Wasylink S, Price LH. A double-blind, placebo-controlled study of risperidone addition in serotonin reuptake inhibitor-refractory obsessive-compulsive disorder. Arch Gen Psychiatry 2000 Aug;57(8):794–801.

McGavin JK, Mann JI, Skeaff CM, Chisholm A. Comparison of a vitamin E-rich diet and supplemental vitamin E on measures of vitamin E status and lipoprotein profile. Eur J Clin Nutr 2001 Jul;55(7):555–61.

Mease PJ, Gladman DD, Keystone EC. Alefacept in combination with methotrexate for the treatment of psoriatic arthritis: results of a randomized, double-blind, placebo-controlled study. Arthritis Rheum 2006 May;54(5):1638–45.

Mease PJ, Gladman DD, Ritchlin CT, Ruderman EM, Steinfeld SD, Choy EH, et al. Adalimumab for the treatment of patients with moderately to severely active psoriatic arthritis: results of a double-blind, randomized, placebo-controlled trial. Arthritis Rheum 2005 Oct;52(10):3279–89.

Miles JM, Leiter L, Hollander P, Wadden T, Anderson JW, Doyle M, et al. Effect of orlistat in overweight and obese patients with type 2 diabetes treated with metformin. Diabetes Care 2002 Jul;25(7):1123–8.

Mori TA, Vandongen R, Beilin LJ, Burke V, Morris J, Ritchie J. Effects of varying dietary fat, fish, and fish oils on blood lipids in a randomized controlled trial in men at risk of heart disease. Am J Clin Nutr 1994 May;59(5):1060–8.

Mottram P, Shige H, Nestel P. Vitamin E improves arterial compliance in middle-aged men and women. Atherosclerosis 1999 Aug;145(2):399–404.

Muchmore DB, Springer J, Miller M. Self-monitoring of blood glucose in overweight type 2 diabetic patients. Acta Diabetol 1994 Dec;31(4):215–9.

Muscettola G, Galzenati M, Balbi A. SAMe versus placebo: a double blind comparison in major depressive disorders. Adv Biochem Psychopharmacol 1982;32:151–6.

Nilsen DW, Albrektsen G, Landmark K, Moen S, Aarsland T, Woie L. Effects of a high-dose concentrate of n-3 fatty acids or corn oil introduced early after an acute myocardial infarction on serum triacylglycerol and HDL cholesterol. Am J Clin Nutr 2001 Jul;74(1):50–6.

Osterud B, Elvevoll E, Barstad H, Brox J, Halvorsen H, Lia K, et al. Effect of marine oils supplementation on coagulation and cellular activation in whole blood. Lipids 1995 Dec;30(12):1111–8.

Paolisso G, Gambardella A, Giugliano D, Galzerano D, Amato L, Volpe C, et al. Chronic intake of pharmacological doses of vitamin E might be useful in the therapy of elderly patients with coronary heart disease. Am J Clin Nutr 1995 Apr;61(4):848–52.

Petrella RJ, DiSilvestro MD, Hildebrand C. Effects of hyaluronate sodium on pain and physical functioning in osteoarthritis of the knee: a randomized, double-blind, placebo-controlled clinical trial. Arch Intern Med 2002 Feb 11;162(3):292–8.

Porkkala-Sarataho EK, Nyyssonen MK, Kaikkonen JE, Poulsen HE, Hayn EM, Salonen RM, et al. A randomized, single-blind, placebo-controlled trial of the effects of 200 mg alpha-tocopherol on the oxidation resistance of atherogenic lipoproteins. Am J Clin Nutr 1998 Nov;68(5):1034–41.

Puhl W, Bernau A, Greiling H, Kopcke W, Pforringer W, Steck KJ, et al. Intra-articular sodium hyaluronate in osteoarthritis of the knee: a multicenter, double-blind study. Osteoarthritis Cartilage 1993 Oct;1(4):233–41.

Raskind MA, Peskind ER, Wessel T, Yuan W. Galantamine in AD: A 6-month randomized, placebo-controlled trial with a 6-month extension. The Galantamine USA-1 Study Group. Neurology 2000 Jun 27;54(12):2261–8.

Raz I, Soskolne V, Stein P. Influence of small-group education sessions on glucose homeostasis in NIDDM. Diabetes Care 1988 Jan;11(1):67–71.

Reaven G, Segal K, Hauptman J, Boldrin M, Lucas C. Effect of orlistat-assisted weight loss in decreasing coronary heart disease risk in patients with syndrome X. Am J Cardiol 2001 Apr 1;87(7):827–31.

Ridgeway NA, Harvill DR, Harvill LM, Falin TM, Forester GM, Gose OD. Improved control of type 2 diabetes mellitus: a practical education/behavior modification program in a primary care clinic. South Med J 1999 Jul;92(7):667–72.

Rockwood K, Fay S, Song X, MacKnight C, Gorman M. Attainment of treatment goals by people with Alzheimer's disease receiving galantamine: a randomized controlled trial. CMAJ 2006 Apr 11;174(8):1099–1105.

Rockwood K, Mintzer J, Truyen L, Wessel T, Wilkinson D. Effects of a flexible galantamine dose in Alzheimer's disease: a randomised, controlled trial. J Neurol Neurosurg Psychiatry 2001 Nov;71(5):589–95.

Rogers SL, Doody RS, Mohs R. Donepezil improves cognition and global function in Alzheimer's disease: a 15-week, double-blind, placebo-controlled study. Donepezil Study Group. Arch Intern Med 1998;158:1021–31.

Rogers SL, Farlow MR, Doody RS, Mohs R, Friedhoff LT. A 24-week, double-blind, placebo-controlled trial of donepezil in patients with Alzheimer's disease. Donepezil Study Group. Neurology 1998 Jan;50(1):136–45.

Rogers SL, Friedhoff LT. The efficacy and safety of donepezil in patients with Alzheimer's disease: results of a US Multicentre, Randomized, Double-Blind, Placebo-Controlled Trial. The Donepezil Study Group. Dementia 1996 Nov-Dec;7(6):293–303.

Rosler M, Anand R, Cicin-Sain A, Gauthier S, Agid Y, Dal-Bianco P, et al. Efficacy and safety of rivastigmine in patients with Alzheimer's disease: international randomised controlled trial. BMJ 1999 Mar 6;318(7184):633–8.

Rossner S, Sjostrom L, Noack R, Meinders A, Noseda G. Weight loss, weight maintenance, and improved cardiovascular risk factors after 2 years treatment with orlistat for obesity. European Orlistat Obesity Study Group. Obes Res 2000 Jan;8(1):49–61.

Rutten G, van Eijk J, de Nobel E, Beek M, van der Velden H. Feasibility and effects of a diabetes type II protocol with blood glucose self-monitoring in general practice. Fam Pract 1990 Dec;7(4):273–8.

Sachdeo RC, Glauser TA, Ritter F, Reife R, Lim P, Pledger G. A double-blind, randomized trial of topiramate in Lennox-Gastaut syndrome. Topiramate YL Study Group. Neurology 1999 Jun 10;52(9):1882–7.

Sachdeo RC, Leroy RF, Krauss GL, Drake ME, Jr., Green PM, Leppik IE, et al. Tiagabine therapy for complex partial seizures. A dose-frequency study. The Tiagabine Study Group. Arch Neurol 1997 May;54(5):595–601.

Sacks FM, Hebert P, Appel LJ, Borhani NO, Applegate WB, Cohen JD, et al. Short report: the effect of fish oil on blood pressure and high-density lipoprotein-cholesterol levels in phase I of the Trials of Hypertension Prevention. J Hypertens 1994 Feb;12(2):209–13.

Sala SF, Miguel RE. Intra-articular hyaluronic acid in the treatment of osteoarthritis of the knee: a short-term study. Eur J Rheumatol Inflamm 1995;15:33–8.

Salmaggi P, Bressa GM, Nicchia G, Coniglio M, La Greca P, Le Grazie C. Double-blind, placebo-controlled study of S-adenosyl-L-methionine in depressed postmenopausal women. Psychother Psychosom 1993;59(1):34–40.

Scale D, Wobig M, Wolpert W. Viscosupplementation of osteoarthritis knees with hylan: a treatment schedule study. Curr Therapeut Res 1994;55:220–32.

Schmidt D, Jacob R, Loiseau P, Deisenhammer E, Klinger D, Despland A, et al. Zonisamide for add-on treatment of refractory partial epilepsy: a European double-blind trial. Epilepsy Res 1993 May;15(1):67–73.

Schwedes U, Siebolds M, Mertes G. Meal-related structured self-monitoring of blood glucose: effect on diabetes control in non-insulin-treated type 2 diabetic patients. Diabetes Care 2002 Nov;25(11):1928–32.

Seltzer B, Zolnouni P, Nunez M, Goldman R, Kumar D, Ieni J, et al. Efficacy of donepezil in early-stage Alzheimer disease: a randomized placebo-controlled trial. Arch Neurol 2004 Dec;61(12):1852–6.

Shapira NA, Ward HE, Mandoki M, Murphy TK, Yang MC, Blier P, et al. A double-blind, placebo-controlled trial of olanzapine addition in fluoxetine-refractory obsessive-compulsive disorder. Biol Psychiatry 2004 Mar 1;55(5):553–5.

Sharief M, Viteri C, Ben-Menachem E, Weber M, Reife R, Pledger G, et al. Double-blind, placebo-controlled study of topiramate in patients with refractory partial epilepsy. Epilepsy Res 1996 Nov;25(3):217–24.

Simmons D, Gamble GD, Foote S, Cole DR, Coster G. The New Zealand Diabetes Passport Study: a randomized controlled trial of the impact of a diabetes passport on risk factors for diabetes-related complications. Diabet Med 2004 Mar;21(3):214–7.

Sivenius J, Kalviainen R, Ylinen A, Riekkinen P. Double-blind study of Gabapentin in the treatment of partial seizures. Epilepsia 1991 Jul-Aug;32(4):539–42.

Sjostrom L, Rissanen A, Andersen T, Boldrin M, Golay A, Koppeschaar H, et al. Randomised placebo-controlled trial of orlistat for weight loss and prevention of weight regain in obese patients. European Multicentre Orlistat Study Group. Lancet 1998 Jul 18;352(9123):167–72.

Stampfer MJ, Willett W, Castelli WP, Taylor JO, Fine J, Hennekens CH. Effect of vitamin E on lipids. Am J Clin Pathol 1983 Jun;79(6):714–6.

Tamir E, Robinson D, Koren R, Agar G, Halperin N. Intra-articular hyaluronan injections for the treatment of osteoarthritis of the knee: a randomized, double blind, placebo controlled study. Clin Exp Rheumatol 2001 May-Jun;19(3):265–70.

Tariot PN, Solomon PR, Morris JC, Kershaw P, Lilienfeld S, Ding C. A 5-month, randomized, placebo-controlled trial of galantamine in AD. The Galantamine USA-10 Study Group. Neurology 2000 Jun 27;54(12):2269–76.

Tassinari CA, Michelucci R, Chauvel P, Chodkiewicz J, Shorvon S, Henriksen O, et al. Double-blind, placebo-controlled trial of topiramate (600 mg daily) for the treatment of refractory partial epilepsy. Epilepsia 1996 Aug;37(8):763–8.

Thomas CS, Bottiglieri T, Edeh J, Carney MW, Reynolds EH, Toone BK. The influence of S-adenosylmethionine (SAM) on prolactin in depressed patients. Int Clin Psychopharmacol 1987 Apr;2(2):97–102.

Tu KS, McDaniel G, Gay JT. Diabetes self-care knowledge, behaviors, and metabolic control of older adults--the effect of a posteducational follow-up program. Diabetes Educ 1993 Jan-Feb;19(1):25–30.

Uthman BM, Rowan AJ, Ahmann PA, Leppik IE, Schachter SC, Sommerville KW, et al. Tiagabine for complex partial seizures: a randomized, add-on, dose-response trial. Arch Neurol 1998 Jan;55(1):56–62.

Vanninen E, Uusitupa M, Siitonen O, Laitinen J, Lansimies E. Habitual physical activity, aerobic capacity and metabolic control in patients with newly-diagnosed type 2 (non-insulin-dependent) diabetes mellitus: effect of 1-year diet and exercise intervention. Diabetologia 1992 Apr;35(4):340–6.

Vrtovec M, Vrtovec B, Briski A, Kocijancic A, Anderson RA, Radovancevic B. Chromium supplementation shortens QTc interval duration in patients with type 2 diabetes mellitus. Am Heart J 2005 Apr;149(4):632–6.

Weinberger M, Kirkman MS, Samsa GP, Shortliffe EA, Landsman PB, Cowper PA, et al. A nurse-coordinated intervention for primary care patients with non-insulin-dependent diabetes mellitus: impact on glycemic control and health-related quality of life. J Gen Intern Med 1995 Feb;10(2):59–66.

White N, Carnahan J, Nugent CA, Iwaoka T, Dodson MA. Management of obese patients with diabetes mellitus: comparison of advice education with group management. Diabetes Care 1986 Sep-Oct;9(5):490–6.

Wilcock GK, Lilienfeld S, Gaens E. Efficacy and safety of galantamine in patients with mild to moderate Alzheimer's disease: multicentre randomised controlled trial. Galantamine International-1 Study Group. BMJ 2000 Dec 9;321(7274):1445–9.

Wilkinson D, Murray J. Galantamine: a randomized, double-blind, dose comparison in patients with Alzheimer's disease. Int J Geriatr Psychiatry 2001 Sep;16(9):852–7.

Willmore LJ, Shu V, Wallin B. Efficacy and safety of add-on divalproex sodium in the treatment of complex partial seizures. The M88-194 Study Group. Neurology 1996 Jan;46(1):49–53.

Wing RR, Epstein LH, Nowalk MP, Scott N, Koeske R, Hagg S. Does self-monitoring of blood glucose levels improve dietary compliance for obese patients with type II diabetes? Am J Med 1986 Nov;81(5):830–6.

Wobig M, Dickhut A, Maier R, Vetter G. Viscosupplementation with hylan G-F 20: a 26-week controlled trial of efficacy and safety in the osteoarthritic knee. Clin Ther 1998 May-Jun;20(3):410–23.

Appendix C. References Dataset 3: Published “Pro-bias” Dataset, 100 Trials

[no author] A randomized trial of aspirin and sulfinpyrazone in threatened stroke. The Canadian Cooperative Study Group. N Engl J Med 1978 Jul 13;299(2):53–9.

Baker RN. Anticoagulant therapy in cerebral infarction. Report on cooperative study. Neurology 1962 Dec;12:823–35.

Baker RN SW, Rose AS. Transient ischemic strokes. A Report of a study (cooperative) of anticoagulant therapy. Neurology 1966;16:841–7.

Baldi F, Bianchi Porro G, Dobrilla G, Iascone C, Lobello R, Marzio L, et al. Cisapride versus placebo in reflux esophagitis. A multicenter double-blind trial. J Clin Gastroenterol 1988 Dec;10(6):614–8.

Balzer K, Bechara G, Bisler H, Clevert HD, Diehm C, Heisig G, et al. Reduction of ischaemic rest pain in advanced peripheral arterial occlusive disease. A double blind placebo controlled trial with iloprost. Int Angiol 1991 Oct-Dec;10(4):229–32.

Bjork K. The efficacy of zimeldine in preventing depressive episodes in recurrent major depressive disorders--a double-blind placebo-controlled study. Acta Psychiatr Scand Suppl 1983;308:182–9.

Bliss WD, Campbell WB. Treatment of limb threatening ischaemia with intravenous iloprost: a randomised double-blind placebo controlled study. U.K. Severe Limb Ischaemia Study Group. Eur J Vasc Surg 1991 Oct;5(5):511–6.

Bradshaw P, Brennan S. Trial of long-term anticoagulant therapy in the treatment of small stroke associated with a normal carotid arteriogram. J Neurol Neurosurg Psychiatry 1975 Jul;38(7):642–7.

Breen KJ, Desmond PV, Whelan G. Treatment of reflux oesophagitis. A randomized, controlled evaluation of cimetidine. Med J Aust 1983 Nov 26;2(11):555–8.

Brock FE, Abri O, Baitsch G, Bechara G, Beck K, Corovic D, et al. [Iloprost in the treatment of ischemic tissue lesions in diabetics. Results of a placebo-controlled multicenter study with a stable prostacyclin derivative]. Schweiz Med Wochenschr 1990 Oct 6;120(40):1477–82.

Brown P. Cimetidine in the treatment of reflux oesophagitis. Med J Aust 1979 Jul 28;2(2):96–7.

Burling TA BG, Robinson JC, Mead AM. Smoking during pregnancy: reduction via objective assessment and directive advice. Behav Ther 1991;22:31–40.

Campieri M, De Franchis R, Bianchi Porro G, Ranzi T, Brunetti G, Barbara L. Mesalazine (5-aminosalicylic acid) suppositories in the treatment of ulcerative proctitis or distal proctosigmoiditis. A randomized controlled trial. Scand J Gastroenterol 1990 Jul;25(7):663–8.

Campieri M, Gionchetti P, Belluzzi A, Brignola C, Tampieri M, Iannone P, et al. Topical treatment with 5-aminosalicylic in distal ulcerative colitis by using a new suppository preparation. A double-blind placebo controlled trial. Int J Colorectal Dis 1990 May;5(2):79–81.

Campieri M, Gionchetti P, Belluzzi A, Brignola C, Tampieri M, Iannone P, et al. Optimum dosage of 5-aminosalicylic acid as rectal enemas in patients with active ulcerative colitis. Gut 1991 Aug;32(8):929–31.

Campieri M GP, Belluzzi A, et al. Sucra lfate, 5amlnosallcyllc acid, and placebo enemas In the treatment of distal ulcerative colltis. Eur J Gastroenterol Hepatol 1991;3:41–4.

Carling L, Cronstedt J, Engqvist A, Kagevi I, Nystrom B, Svedberg LE, et al. Sucralfate versus placebo in reflux esophagitis. A double-blind multicenter study. Scand J Gastroenterol 1988 Nov;23(9):1117–24.

Coppen A, Ghose K, Montgomery S, Rama Rao VA, Bailey J, Jorgensen A. Continuation therapy with amitriptyline in depression. Br J Psychiatry 1978 Jul;133:28–33.

Dew MJ, Hughes P, Harries AD, Williams G, Evans BK, Rhodes J. Maintenance of remission in ulcerative colitis with oral preparation of 5-aminosalicylic acid. Br Med J (Clin Res Ed) 1982 Oct 9;285(6347):1012.

Diehm C, Abri O, Baitsch G, Bechara G, Beck K, Breddin HK, et al. [Iloprost, a stable prostacyclin derivative, in stage 4 arterial occlusive disease. A placebo-controlled multicenter study]. Dtsch Med Wochenschr 1989 May 19;114(20):783–8.

Duke RJ, Bloch RF, Turpie AG, Trebilcock R, Bayer N. Intravenous heparin for the prevention of stroke progression in acute partial stable stroke. Ann Intern Med 1986 Dec;105(6):825–8.

Duke RJ TA, Bloch RF, Trebilcock RG. Clinical trial of low-dose subcutaneous heparin for the prevention of stroke progression: natural history of acute partial stroke and stroke-in-evolution. In: Reivich M HH, ed. Cerebrovascular Disease. New York: Raven Press; 1983. p. 399–405.

Elias A, Milandre L, Lagrange G, Aillaud MF, Alonzo B, Toulemonde F, et al. [Prevention of deep venous thrombosis of the leg by a very low molecular weight heparin fraction (CY 222) in patients with hemiplegia following cerebral infarction: a randomized pilot study (30 patients)]. Rev Med Interne 1990 Jan-Feb;11(1):95–8.

Ershoff DH, Mullen PD, Quinn VP. A randomized trial of a serialized self-help smoking cessation program for pregnant women in an HMO. Am J Public Health 1989 Feb;79(2):182–7.

Falloon IR, Boyd JL, McGill CW, Razani J, Moss HB, Gilderman AM. Family management in the prevention of exacerbations of schizophrenia: a controlled study. N Engl J Med 1982 Jun 17;306(24):1437–40.

Farup PG, Weberg R, Berstad A, Wetterhus S, Dahlberg O, Dybdahl J, et al. Low-dose antacids versus 400 mg cimetidine twice daily for reflux oesophagitis. A comparative, placebo-controlled, multicentre study. Scand J Gastroenterol 1990 Mar;25(3):315–20.

Festen HP, Driessen WM, Lamers CB, Van Tongeren JH. Cimetidine in the treatment of severe ulcerative reflux oesophagitis; results of an 8-week double-blind study and of subsequent long-term maintenance treatment. Neth J Med 1980;23(6):237–40.

Feurle GE, Theuer D, Velasco S, Barry BA, Wordehoff D, Sommer A, et al. Olsalazine versus placebo in the treatment of mild to moderate ulcerative colitis: a randomised double blind trial. Gut 1989 Oct;30(10):1354–61.

Fiasse R, Hanin C, Lepot A, Descamps C, Lamy F, Dive C. Controlled trial of cimetidine in reflux esophagitis. Dig Dis Sci 1980 Oct;25(10):750–5.

Fields WS, Lemak NA, Frankowski RF, Hardy RJ. Controlled trial of aspirin in cerebral ischemia. Stroke 1977 May-Jun;8(3):301–14.

Fields WS, Lemak NA, Frankowski RF, Hardy RJ. Controlled trial of aspirin in cerebral ischemia. Part II: surgical group. Stroke 1978 Jul-Aug;9(4):309–19.

Gielen. Personal Communication. [Personal Communication]. In press 1993.

Gionchetti P, Campieri M, Belluzzi A, Brignola C, Tampieri M, Iannone P, et al. Pentasa in maintenance treatment of ulcerative colitis. Gastroenterology 1990 Jan;98(1):251.

Glen AI, Johnson AL, Shepherd M. Continuation therapy with lithium and amitriptyline in unipolar depressive illness: a randomized, double-blind, controlled trial. Psychol Med 1984 Feb;14(1):37–50.

Goy JA, Maynard JH, McNaughton WM, O'Shea A. Ranitidine and placebo in the treatment of reflux oesophagitis. A double-blind randomized trial. Med J Aust 1983 Nov 26;2(11):558–61.

Guilmot J-L DE. Treatment of lower limb ischaemia due to atherosclerosis in diabetic and non-diabetic patients with Iloprost, a stable analogue of prostacyclin: results of the French multicentre trial. Drug Investigation 1991;3(5):351–9.

Hakim AM, Furlan AJ, Hart RG. Immediate anticoagulation of embolic stroke: a randomized trial. Cerebral Embolism Study Group. Stroke 1983;14(5):668–76.

Hanauer S, Schwartz J, Roufall W, Robinson M, Cello J, Safdi M, et al. Dose-ranging study of oral mesalamine capsule (PENATASA) for active ulcerative colitis. Gastroenterology [Abstract] 1989;96:A195.

Harrison RF, O'Moore RR, McSweeney J. Idiopathic infertility: a trial of bromocriptine versus placebo. Ir Med J 1979 Nov 30;72(11):479–82.

Hetzel DJ, Dent J, Reed WD, Narielvala FM, Mackinnon M, McCarthy JH, et al. Healing and relapse of severe peptic esophagitis after treatment with omeprazole. Gastroenterology 1988 Oct;95(4):903–12.

Hetzel DJ, Shearman DJ, Bochner F, Imhoff DM, Gibson GE, Fitch RJ, et al. Azodisalicylate (olsalazine) in the treatment of active ulcerative colitis. A placebo controlled clinical trial and assessment of drug disposition. . J Gastroent Hepatol 1986;1:257–66.

Hjalmarson AI, Hahn L, Svanberg B. Stopping smoking in pregnancy: effect of a self-help manual in controlled trial. Br J Obstet Gynaecol 1991 Mar;98(3):260–4.

Hogarty GE, Anderson CM, Reiss DJ, Kornblith SJ, Greenwald DP, Javna CD, et al. Family psychoeducation, social skills training, and maintenance chemotherapy in the aftercare treatment of schizophrenia. I. One-year effects of a controlled study on relapse and expressed emotion. Arch Gen Psychiatry 1986 Jul;43(7):633–42.

Hull RD, Raskob GE, Pineo GF, Green D, Trowbridge AA, Elliott CG, et al. Subcutaneous low-molecular-weight heparin compared with continuous intravenous heparin in the treatment of proximal-vein thrombosis. N Engl J Med 1992 Apr 9;326(15):975–82.

Ireland A, Mason CH, Jewell DP. Controlled trial comparing olsalazine and sulphasalazine for the maintenance treatment of ulcerative colitis. Gut 1988 Jun;29(6):835–7.

Jestico J, Harrison MJ, Marshall J. Trial of aspirin during weaning patients with transient ischaemic attacks from anticoagulants. Br Med J 1978 May 6;1(6121):1188.

Johansson KE, Boeryd B, Johansson K, Tibbling L. Double-blind crossover study of ranitidine and placebo in gastro-oesophageal reflux disease. Scand J Gastroenterol 1986 Sep;21(7):769–78.

Kane JM, Quitkin FM, Rifkin A, Ramos-Lorenzi JR, Nayak DD, Howard A. Lithium carbonate and imipramine in the prophylaxis of unipolar and bipolar II illness: a prospective, placebo-controlled comparison. Arch Gen Psychiatry 1982 Sep;39(9):1065–9.

Kiilerich S, Ladefoged K, Rannem T, Ranlov PJ. Prophylactic effects of olsalazine v sulphasalazine during 12 months maintenance treatment of ulcerative colitis. The Danish Olsalazine Study Group. Gut 1992 Feb;33(2):252–5.

Klerman GL, Dimascio A, Weissman M, Prusoff B, Paykel ES. Treatment of depression by drugs and psychotherapy. Am J Psychiatry 1974 Feb;131(2):186–91.

Leff J, Kuipers L, Berkowitz R, Sturgeon D. A controlled trial of social intervention in the families of schizophrenic patients: two year follow-up. Br J Psychiatry 1985 Jun;146:594–600.

Lehtola J, Niemela S, Martikainen J, Krekela I. Ranitidine, 150 mg three times a day, in the treatment of reflux oesophagitis. A placebo-controlled, double-blind study. Scand J Gastroenterol 1986 Mar;21(2):175–80.

Lepoutre L, Van der Spek P, Vanderlinden I, Bollen J, Laukens P. Healing of grade-II and III oesophagitis through motility stimulation with cisapride. Digestion 1990;45(2):109–14.

Lepsien G, Sonnenberg A, Berges W, Weber KB, Wienbeck M, Siewert JR, et al. [Treatment of reflux oesophagitis with cimetidine (author's transl)]. Dtsch Med Wochenschr 1979 Jun 22;104(25):901–6.

Lindmarker P, Holmstrom M, Granqvist S, Johnsson H, Lockner D. Comparison of once-daily subcutaneous Fragmin with continuous intravenous unfractionated heparin in the treatment of deep vein thrombosis. Thromb Haemost 1994 Aug;72(2):186–90.

Lopaciuk S, Meissner AJ, Filipecki S, Zawilska K, Sowier J, Ciesielski L, et al. Subcutaneous low molecular weight heparin versus subcutaneous unfractionated heparin in the treatment of deep vein thrombosis: a Polish multicenter trial. Thromb Haemost 1992 Jul 6;68(1):14–8.

Marshall J, Shaw DA. Anticoagulant therapy in acute cerebrovascular accidents. A controlled trial. Lancet 1960 May 7;1(7132):995–8.

McBain JC, Pepperell RJ. Use of bromocriptine in unexplained infertility. Clin Reprod Fertil 1982 Jun;1(2):145–50.

McCallum RW, Fink SM, Winnan GR, Avella J, Callachan C. Metoclopramide in gastroesophageal reflux disease: rationale for its use and results of a double-blind trial. Am J Gastroenterol 1984 Mar;79(3):165–72.

McIntyre PB, Rodrigues CA, Lennard-Jones JE, Barrison IG, Walker JG, Baron JH, et al. Balsalazide in the maintenance treatment of patients with ulcerative colitis, a double-blind comparison with sulphasalazine. Aliment Pharmacol Ther 1988 Jun;2(3):237–43.

Mindham RH, Howland C, Shepherd M. An evaluation of continuation therapy with tricyclic antidepressants in depressive illness. Psychol Med 1973 Feb;3(1):5–17.

Montgomery SA, Dufour H, Brion S, Gailledreau J, Laqueille X, Ferrey G, et al. The prophylactic efficacy of fluoxetine in unipolar depression. Br J Psychiatry Suppl 1988 Sep(3):69–76.

Mulder CJ, Tytgat GN, Weterman IT, Dekker W, Blok P, Schrijver M, et al. Double-blind comparison of slow-release 5-aminosalicylate and sulfasalazine in remission maintenance in ulcerative colitis. Gastroenterology 1988 Dec;95(6):1449–53.

Niemela S, Jaaskelainen T, Lehtola J, Martikainen J, Krekela I, Sarna S, et al. Pirenzepine in the treatment of reflux oesophagitis. A placebo-controlled, double-blind study. Scand J Gastroenterol 1986 Dec;21(10):1193–9.

Norgren L, Alwmark A, Angqvist KA, Hedberg B, Bergqvist D, Takolander R, et al. A stable prostacyclin analogue (iloprost) in the treatment of ischaemic ulcers of the lower limb. A Scandinavian-Polish placebo controlled, randomised multicenter study. Eur J Vasc Surg 1990 Oct;4(5):463–7.

O'Connor AM, Davies BL, Dulberg CS, Buhler PL, Nadon C, McBride BH, et al. Effectiveness of a pregnancy smoking cessation program. J Obstet Gynecol Neonatal Nurs 1992 Sep-Oct;21(5):385–92.

Palmer RH, Frank WO, Rockhold FW, Wetherington JD, Young MD. Cimetidine 800 mg twice daily for healing erosions and ulcers in gastroesophageal reflux disease. J Clin Gastroenterol 1990;12 Suppl 2:S29–34.

Pearce JM, Gubbay SS, Walton JN. Long-Term Anticoagulant Therapy in Transient Cerebral Ischaemic Attacks. Lancet 1965 Jan 2;1(7375):6–9.

Petersen L, Handel J, Kotch J, Podedworny T, Rosen A. Smoking reduction during pregnancy by a program of self-help and clinical support. Obstet Gynecol 1992 Jun;79(6):924–30.

Pince J. Thromboses veineuses des membres inferieurs et embolies pulmonaires au cours des accidents vasculaires cerebraux. A propos d'un essai comparitif de traitement preventif (These pour le doctorat d'etat en medecine). In press 1981.

Prandoni P, Lensing AW, Buller HR, Carta M, Cogo A, Vigo M, et al. Comparison of subcutaneous low-molecular-weight heparin with intravenous standard heparin in proximal deep-vein thrombosis. Lancet 1992 Feb 22;339(8791):441–5.

Price JH, Krol RA, Desmond SM, Losh DP, Roberts SM, Snyder FF. Comparison of three antismoking interventions among pregnant women in an urban setting: a randomized trial. Psychol Rep 1991 Apr;68(2):595–604.

Prins MH, Gelsema R, Sing AK, van Heerde LR, den Ottolander GJ. Prophylaxis of deep venous thrombosis with a low-molecular-weight heparin (Kabi 2165/Fragmin) in stroke patients. Haemostasis 1989;19(5):245–50.

Quik RF, Cooper MJ, Gleeson M, Hentschel E, Schuetze K, Kingston RD, et al. A comparison of two doses of nizatidine versus placebo in the treatment of reflux oesophagitis. Aliment Pharmacol Ther 1990 Apr;4(2):201–11.

Reuther R. DW. Aspirin in patients with cerebral ischaemia and normal angiograms or non-surgical lesions. Acetylsalicylic Acid in Cerebral Ischaemia and Coronary Heart Disease 1978:97–106.

Rijk MC, van Lier HJ, van Tongeren JH. Relapse-preventing effect and safety of sulfasalazine and olsalazine in patients with ulcerative colitis in remission: a prospective, double-blind, randomized multicenter study. The Ulcerative Colitis Multicenter Study Group. Am J Gastroenterol 1992 Apr;87(4):438–42.

Riley SA, Mani V, Goodman MJ, Herd ME, Dutt S, Turnberg LA. Comparison of delayed-release 5-aminosalicylic acid (mesalazine) and sulfasalazine as maintenance treatment for patients with ulcerative colitis. Gastroenterology 1988 Jun;94(6):1383–9.

Rutgeerts P. Comparative efficacy of coated, oral 5-aminosalicylic acid (Claversal) and sulphasalazine for maintaining remission of ulcerative colitis. International Study Group. Aliment Pharmacol Ther 1989 Apr;3(2):183–91.

Sandset PM, Dahl T, Stiris M, Rostad B, Scheel B, Abildgaard U. A double-blind and randomized placebo-controlled trial of low molecular weight heparin once daily to prevent deep-vein thrombosis in acute ischemic stroke. Semin Thromb Hemost 1990 Oct;16 Suppl:25–33.

Schroeder KW, Tremaine WJ, Ilstrup DM. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. A randomized study. N Engl J Med 1987 Dec 24;317(26):1625–9.

Seager CP, Bird RL. Imipramine with electrical treatment in depression--a controlled trial. J Ment Sci 1962 Sep;108:704–7.

Secker-Walker RH, Solomon LJ, Flynn BS, Skelly JM, Lepage SS, Goodwin GD, et al. Individualized smoking cessation counseling during prenatal and early postnatal care. Am J Obstet Gynecol– 1994 Nov;171(5):1347–55.

Sexton M, Hebel JR. A clinical trial of change in maternal smoking and its effect on birth weight. JAMA 1984 Feb 17;251(7):911–5.

Sherbaniuk R, Wensel R, Bailey R, Trautman A, Grace M, Kirdeikis P, et al. Ranitidine in the treatment of symptomatic gastroesophageal reflux disease. J Clin Gastroenterol 1984 Feb;6(1):9–15.

Simonneau G, Charbonnier B, Decousus H, Planchon B, Ninet J, Sie P, et al. Subcutaneous low-molecular-weight heparin compared with continuous intravenous unfractionated heparin in the treatment of proximal deep vein thrombosis. Arch Intern Med 1993 Jul 12;153(13):1541–6.

Sninsky CA, Cort DH, Shanahan F, Powers BJ, Sessions JT, Pruitt RE, et al. Oral mesalamine (Asacol) for mildly to moderately active ulcerative colitis. A multicenter study. Ann Intern Med 1991 Sep 1;115(5):350–5.

Sontag S, Robinson M, McCallum RW, Barwick KW, Nardi R. Ranitidine therapy for gastroesophageal reflux disease. Results of a large double-blind trial. Arch Intern Med 1987 Aug;147(8):1485–91.

Sorensen PS, Pedersen H, Marquardsen J, Petersson H, Heltberg A, Simonsen N, et al. Acetylsalicylic acid in the prevention of stroke in patients with reversible cerebral ischemic attacks. A Danish cooperative study. Stroke 1983 Jan-Feb;14(1):15–22.

Stein MK, Rickels K, Weise CC. Maintenance therapy with amitriptyline: a controlled trial. Am J Psychiatry 1980 Mar;137(3):370–1.

Sutherland LR, Martin F, Greer S, Robinson M, Greenberger N, Saibil F, et al. 5-Aminosalicylic acid enema in the treatment of distal ulcerative colitis, proctosigmoiditis, and proctitis. Gastroenterology 1987 Jun;92(6):1894–8.

Sutherland LR, Robinson M, Onstad G, Peppercorn M, Greenberger N, Goodman M, et al. A double-blind, placebo controlled, multicentre study of the efficacy and safety of 5-aminosalicylic acid tablets in the treatment of ulcerative colitis. Can J Gastroenterol 1990;4:463–7.

Tarrier N, Barrowclough C, Vaughn C, Bamrah JS, Porceddu K, Watts S, et al. The community management of schizophrenia. A controlled trial of a behavioural intervention with families to reduce relapse. Br J Psychiatry 1988 Oct;153:532–42.

Turpie AG, Levine MN, Hirsh J, Carter CJ, Jay RM, Powers PJ, et al. Double-blind randomised trial of Org 10172 low-molecular-weight heparinoid in prevention of deep-vein thrombosis in thrombotic stroke. Lancet 1987 Mar 7;1(8532):523–6.

Vaughan K, Doyle M, McConaghy N, Blaszczynski A, Fox A, Tarrier N. The Sydney intervention trial: a controlled trial of relatives' counselling to reduce schizophrenic relapse. Soc Psychiatry Psychiatr Epidemiol 1992 Jan;27(1):16–21.

Weiss. Klinische Erfahrungen mit Ulcogant bei der Reflux-oesophagitis. Swiss Med 1983;5:21–4.

Wesdorp E, Bartelsman J, Pape K, Dekker W, Tytgat GN. Oral cimetidine in reflux esophagitis: a double blind controlled trial. Gastroenterology 1978 May;74(5 Pt 1):821–4.

Wesdorp IC, Dekker W, Klinkenberg-Knol EC. Treatment of reflux oesophagitis with ranitidine. Gut 1983 Oct;24(10):921–4.

Windsor RA, Cutter G, Morris J, Reese Y, Manzella B, Bartlett EE, et al. The effectiveness of smoking cessation methods for smokers in public health maternity clinics: a randomized trial. Am J Public Health 1985 Dec;75(12):1389–92.

Windsor RA, Lowe JB, Perkins LL, Smith-Yoder D, Artz L, Crawford M, et al. Health education for pregnant smokers: its behavioral impact and cost benefit. Am J Public Health 1993 Feb;83(2):201–6.

Wright CS, Steele SJ, Jacobs HS. Value of bromocriptine in unexplained primary infertility: a double-blind controlled trial. Br Med J 1979 Apr 21;1(6170):1037–9.

Appendix D. Comparison Fixed-Effects Model Results

Table D1. Difference in effect sizes (EPC reports), FE

Quality Feature	Number Criterion Met	Number Criterion Not Met	Effect Size in Trials With Criterion Met		Effect Size in Trials With Criterion Not Met		Effect Size Difference	
			ES	95% CI	ES	95% CI	ESdiff	95% CI
Randomization adequate	44	121	0.59	(0.56, 0.61)	1.08	(1.07, 1.10)	-0.50	(-0.53, -0.47)*
Allocation concealment	38	127	0.56	(0.53, 0.59)	1.09	(1.07, 1.10)	-0.53	(-0.56, -0.49)*
Similar baseline	100	65	1.00	(0.99, 1.02)	0.87	(0.83, 0.90)	0.14	(0.10, 0.17)*
Assessor blind	157	8	1.00	(0.98, 1.01)	0.41	(0.31, 0.51)	0.59	(0.49, 0.68)*
Care provider blind	120	45	1.09	(1.07, 1.10)	0.14	(0.10, 0.18)	0.95	(0.91, 0.99)*
Patient blind	130	35	1.09	(1.07, 1.10)	0.08	(0.04, 0.12)	1.01	(0.96, 1.05)*
Acceptable dropout rate	96	69	1.15	(1.13, 1.16)	0.24	(0.21, 0.27)	0.90	(0.87, 0.93)*
Original group (ITT)	101	64	1.03	(1.02, 1.04)	0.40	(0.35, 0.45)	0.63	(0.58, 0.68)*
Similar co-interventions	142	23	1.02	(1.00, 1.03)	0.25	(0.18, 0.31)	0.77	(0.71, 0.83)*
Acceptable compliance	79	86	1.18	(1.17, 1.20)	0.29	(0.27, 0.32)	0.89	(0.86, 0.92)*
Similar timing	161	4	0.99	(0.98, 1.01)	0.28	(0.15, 0.42)	0.71	(0.58, 0.84)*

* p<0.05
EPC = Evidence-based Practice Center; FE = based on fixed-effects model; ES = effect size; CI = confidence interval; ESdiff = effect size difference; ITT = intention to treat

Table D2. Comparison of different quality cutoffs using a total score (EPC reports), FE

Cutoff	Number Equal or Above Cut-off	Number Below Cutoff	High Quality		Low Quality		Difference	
			ES	95% CI	ES	95% CI	ESdiff	95% CI
≥9 vs <9	42	123	1.22	(1.21, 1.24)	0.29	(0.26, 0.31)	0.93	(0.91, 0.96)*
≥8 vs <8	65	100	1.16	(1.15, 1.18)	0.27	(0.24, 0.30)	0.89	(0.86, 0.93)*
≥7 vs <7	103	62	1.03	(1.02, 1.04)	0.41	(0.37, 0.46)	0.62	(0.57, 0.67)*
≥6 vs <6	135	30	1.01	(1.00, 1.02)	0.21	(0.14, 0.28)	0.80	(0.73, 0.88)*
≥5 vs <5	149	16	1.00	(0.99, 1.02)	0.10	(0.00, 0.20)	0.90	(0.80, 1.00)*
≥4 vs <4	160	5	0.99	(0.98, 1.00)	0.06	(-0.16, 0.27)	0.93	(0.72, 1.15)

* p<0.05
EPC = Evidence-based Practice Center; FE = based on fixed effects model; ES = effect size; CI = confidence interval; ESdiff = effect size difference

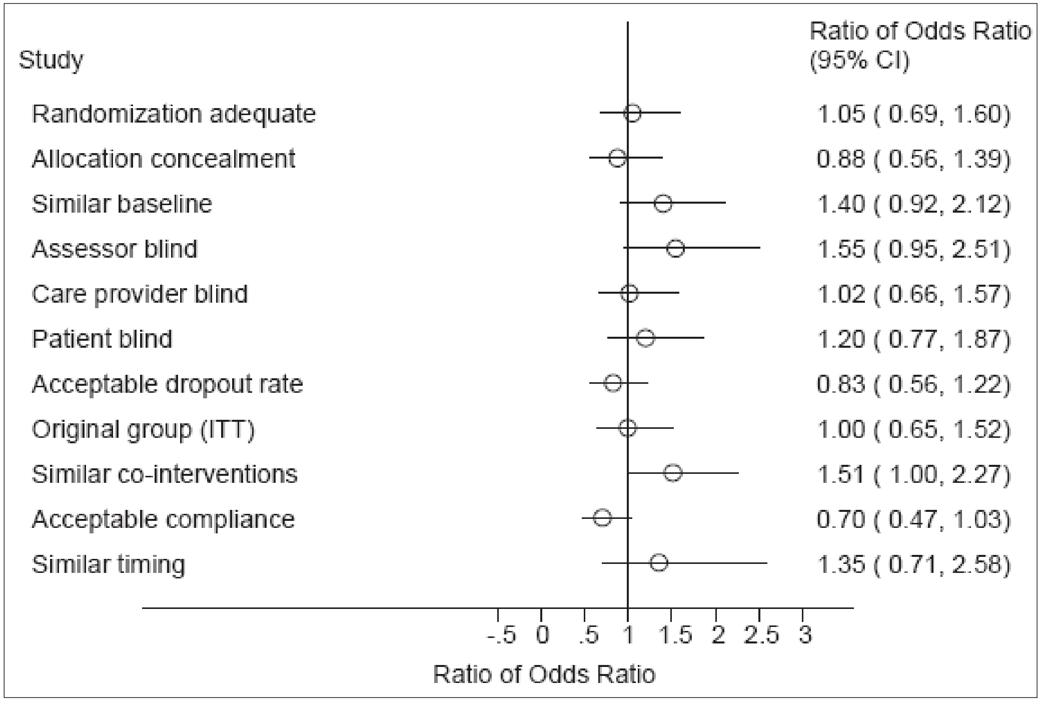
Appendix E. Comparison Random Effects Meta-regression Results

Table E1. Difference in odds ratios for proposed quality criteria ("pro-bias"), R

Quality Feature	Number Criterion Met	Number Criterion Not Met	Effect Size in Trials With Criterion Met		Effect Size in Trials With Criterion Not Met		Effect Size Difference	
			OR	95% CI	OR	95% CI	ROR	95% CI
Randomization adequate	34	66	0.44	(0.31, 0.59)	0.46	(0.37, 0.59)	1.05	(0.69, 1.60)
Allocation concealment	26	74	0.50	(0.34, 0.75)	0.44	(0.35, 0.56)	0.88	(0.56, 1.39)
Similar baseline	36	64	0.37	(0.26, 0.52)	0.51	(0.40, 0.66)	1.40	(0.92, 2.12)
Assessor blind	78	22	0.42	(0.34, 0.52)	0.65	(0.42, 1.00)	1.55	(0.95, 2.51)
Care provider blind	69	31	0.45	(0.36, 0.58)	0.46	(0.32, 0.66)	1.02	(0.66, 1.57)
Patient blind	72	28	0.44	(0.35, 0.55)	0.52	(0.36, 0.76)	1.20	(0.77, 1.87)
Acceptable dropout rate	62	38	0.50	(0.38, 0.64)	0.41	(0.30, 0.55)	0.83	(0.57, 1.22)
Original group (ITT)	29	71	0.46	(0.32, 0.65)	0.46	(0.36, 0.58)	1.00	(0.65, 1.52)
Similar co-interventions	68	32	0.40	(0.32, 0.51)	0.60	(0.43, 0.84)	1.51	(1.00, 2.27)*
Acceptable compliance	46	54	0.56	(0.42, 0.75)	0.39	(0.30, 0.51)	0.70	(0.47, 1.03)
Similar timing	89	11	0.44	(0.36, 0.55)	0.60	(0.32, 1.11)	1.35	(0.71, 2.58)

* p<0.05
R = random effects meta-regression; OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios;
ITT = intention to treat

Figure E1. Differences in effect sizes based on quality criteria (“pro-bias”), random effects meta-regression



CI = confidence interval; ITT = intention to treat

Table E2. Comparison of different quality cutoffs using a total score (“pro-bias”), R

Cut-off	Number Equal or Above Cut-off	Number Below Cut-off	High Quality		Low Quality		Difference	
			OR	95% CI	OR	95% CI	ROR	95% CI
≥9 vs <9	14	86	0.39	(0.23, 0.67)	0.47	(0.38, 0.58)	1.20	(0.68, 2.10)
≥8 vs <8	26	74	0.46	(0.32, 0.67)	0.46	(0.36, 0.58)	0.99	(0.64, 1.55)
≥7 vs <7	44	56	0.39	(0.29, 0.53)	0.51	(0.40, 0.67)	1.31	(0.88, 1.95)
≥6 vs <6	62	38	0.46	(0.36, 0.59)	0.46	(0.32, 0.63)	0.99	(0.66, 1.49)
≥5 vs <5	76	24	0.45	(0.36, 0.57)	0.47	(0.32, 0.71)	1.05	(0.66, 1.67)
≥4 vs <4	86	14	0.46	(0.36, 0.55)	0.53	(0.32, 0.88)	1.19	(0.69, 2.06)

R = random effects meta-regression; OR = odds ratio; CI = confidence interval; ROR = ratio of odds ratios

Appendix F. Quality Rating Form

Article ID: Reviewer:

First Author, Year: Meta-analysis:
(Last Name Only)

CBRG Quality Items

Was the method of randomization adequate?

Yes ....❑

No..❑

Don't know..❑

Was the treatment allocation concealed?

Yes ....❑

No..❑

Don't know..❑

Were the groups similar at baseline regarding the most important prognostic indicators?

Yes ....❑

No..❑

Don't know..❑

Was the outcome assessor blinded?

Yes ....❑

No..❑

Don't know..❑

Was the care provider blinded?

Yes ....❑

No..❑

Don't know..❑

Were patients blinded?

Yes ....❑

No..❑

Don't know..❑

Was the dropout rate described and acceptable?

Yes ....❑

No..❑

Don't know..❑

Were all randomized participants analyzed in the group to which they were originally assigned?

Yes ....❑

No..❑

Don't know..❑

Were co-interventions avoided or similar?

Yes ....❑

No..❑

Don't know..❑

Was the compliance acceptable in all groups?

Yes ....❑

No..❑

Don't know..❑

Was the timing of the outcome assessment similar in all groups?

Yes ....❑

No..❑

Don't know..❑

Scoring Guidelines Cochrane Back Review Group

Randomization sequence

A random (unpredictable) assignment sequence. Examples of adequate methods are coin toss (for studies with two groups), rolling a dice (for studies with two or more groups), drawing of balls of different colours, drawing of ballots with the study group labels from a dark bag, computergenerated random sequence, pre-ordered sealed envelops, sequentially-ordered vials, telephone call to a central office, and pre-ordered list of treatment assignments Examples of inadequate methods are: alternation, birth date, social insurance/security number, date in which they are invited to participate in the study, and hospital registration number

Allocation concealment

Assignment generated by an independent person not responsible for determining the eligibility of the patients. This person has no information about the persons included in the trial and has no influence on the assignment sequence or on the decision about eligibility of the patient.

Patient blinding

This item should be scored “yes” if the index and control groups are indistinguishable for the patients or if the success of blinding was tested among the patients and it was successful.

Care provider blinding

This item should be scored “yes” if the index and control groups are indistinguishable for the care providers or if the success of blinding was tested among the care providers and it was successful

Assessor blinding

Adequacy of blinding should be assessed for the primary outcomes. This item should be scored “yes” if the success of blinding was tested among the outcome assessors and it was successful or:

- **for patient-reported outcomes** in which the patient is the outcome assessor (e.g., pain, disability): the blinding procedure is adequate for outcome assessors if participant blinding is scored “yes”
- **for outcome criteria assessed during scheduled visit and that supposes a contact between participants and outcome assessors** (e.g., clinical examination): the blinding procedure is adequate if patients are blinded, and the treatment or adverse effects of the treatment cannot be noticed during clinical examination
- **for outcome criteria that do not suppose a contact with participants** (e.g., radiography, magnetic resonance imaging): the blinding procedure is adequate if the treatment or adverse effects of the treatment cannot be noticed when assessing the main outcome
- **for outcome criteria that are clinical or therapeutic events** that will be determined by the interaction between patients and care providers (e.g., co-interventions, hospitalization length, treatment failure), in which the care provider is the outcome assessor: the blinding procedure is adequate for outcome assessors if item “E” is scored “yes”
- **for outcome criteria that are assessed from data of the medical forms**: the blinding procedure is adequate if the treatment or adverse effects of the treatment cannot be noticed on the extracted data

Dropouts

The number of participants who were included in the study but did not complete the observation period or were not included in the analysis must be described and reasons given. If the percentage of withdrawals and dropouts does not exceed 20% for short-term follow-up and 30% for long-term followup and does not lead to substantial bias a 'yes' is scored. (N.B. these percentages are arbitrary, not supported by literature).

ITT

All randomized patients are reported/analyzed in the group they were allocated to by randomization for the most important moments of effect measurement (minus missing values) irrespective of noncompliance and co-interventions.

Baseline comparability

In order to receive a “yes”, groups have to be similar at baseline regarding demographic factors, duration and severity of complaints, percentage of patients with neurological symptoms, and value of main outcome measure(s).

Co-Interventions

This item should be scored “yes” if there were no co-interventions or they were similar between the index and control groups.

Compliance

The reviewer determines if the compliance with the interventions is acceptable, based on the reported intensity, duration, number and frequency of sessions for both the index intervention and control intervention(s). For example, physiotherapy treatment is usually administered over several sessions; therefore it is necessary to assess how many sessions each patient attended. For single-session interventions (for ex: surgery), this item is irrelevant.

Timing

Timing of outcome assessment should be identical for all intervention groups and for all important outcome assessments.

Note: These instructions are adapted from van Tulder 2003, Boutron et al, 2005 (CLEAR NPT) and the Cochrane Handbook of Reviews of Interventions2;5;9. 2008 Updated Guidelines for Systematic Reviews 9April 2008

Jadad Scale

Dimension			Sub Score
Randomization	1. Was the study described as randomized (this includes the use of words such as randomly, random, and randomization)? = 1 point	Give 1 additional point if: For question 1, the method to generate the sequence of randomization was described and it was appropriate (table of random numbers, computer generated, etc.) Deduct 1 point if: For question 1, the method to generate the sequence of randomization was described and it was inappropriate (patients were allocated alternately, or according to date of birth, hospital number, etc.)	
Blinding	2. Was the study described as double blind? = 1 point	Give 1 additional point: If for question 2 the method of double blinding was described and it was appropriate (identical placebo, active placebo, dummy, etc.) Deduct 1 point: If for question 2 the study was described as double blind but the method of blinding was inappropriate (e.g., comparison of tablet vs. injection with no double dummy)	
Withdrawals and dropouts	3. Was there a description of withdrawals and dropouts? = 1 point		
		TOTAL JADAD SCORE	

Jadad Guidelines for Assessment

1. Randomization

A method to generate the sequence of randomization will be regarded as appropriate if it allowed each study participant to have the same chance of receiving each intervention and the investigators could not predict which treatment was next. Methods of allocation using date of birth, date of admission, hospital numbers, or alternation should be not regarded as appropriate.

2. Double blinding

A study must be regarded as double blind if the word "double blind" is used. The method will be regarded as appropriate if it is stated that neither the person doing the assessments nor the study participant could identify the intervention being assessed, or if in the absence of such a statement the use of active placebos, identical placebos, or dummies is mentioned.

3. Withdrawals and dropouts

Participants who were included in the study but did not complete the observation period or who were not included in the analysis must be described. The number and the reasons for withdrawal in each group must be stated. If there were no withdrawals, it should be stated in the article. If there is no statement on withdrawals, this item must be given no points.

Schulz's (1995) quality dimensions
(circle appropriate category)

1. Concealment of Treatment Allocation

a) Adequately concealed trial (i.e. central randomization; numbered or coded bottles or containers; drugs prepared by the pharmacy; serially numbered; opaque, sealed envelopes; or other description that contained elements convincing of concealment

b) Inadequately concealed trial (i.e. alternation or reference to case record numbers or dates of birth

c) Unclearly concealed trial (authors did either not report an allocation concealment approach at all or reported an approach that did not fall into the categories above

2. Generation of Allocation Sequence

a) Adequately sequence generation (random-number table, computer random-number generator, coin tossing, or shuffling)

b) Publication does not report one of the adequate approaches, those with inadequate sequence generation

3. Inclusion in the Analysis of All Randomized Participants

a) Publication reports or gives the impression that no exclusions have taken place (often not explicit)

b) Publication reports exclusions (e.g., protocol deviation, withdrawals, dropouts, loss to follow-up)

4. Double Blinding

a) Double-blinding reported

b) Double-blinding not reported